h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACULTY OF COMPUTER SCIENCE

Hochschule Darmstadt, University of Applied Science
– Department of Computer Science –

SPEAKER VERIFICATION USING I-VECTORS

Evaluation of text-independent speaker verification systems based on identity-vectors in short and variant duration scenarios

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

vorgelegt von

ANDREAS NAUTSCH

Referent:	Prof. Dr. Christoph Busch
Korreferent:	Dr. Christian Rathgeb
Firmenbetreuer:	Prof. Dr. Herbert Reininger
Ausgabedatum:	02.10.2013
Abgabedatum:	02.04.2014

atip GmbH
Daimlerstraße 32
60314 Frankfurt am Main

Andreas Nautsch: *Speaker verification using i-vectors,* Evaluation of text-independent speaker verification systems based on identity-vectors in short and variant duration scenarios, Master of Science (M.Sc.), © April 1, 2014

SUPERVISORS:
Prof. Dr. Christoph Busch
Dr. Christian Rathgeb
Prof. Dr. Herbert Reininger

LOCATION:
Darmstadt

TIME FRAME:
02.10.2013 — 02.04.2014

ERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Darmstadt, 1. April 2014

Andreas Nautsch

ABSTRACT

Speaker verification becomes more important as a biometric key security solution in industry, forensic, and governmental terms. Telephone-based authentication concepts ensuring purposes of data privacy get more popular e. g., data encryption on mobile devices, or user validations on contact-centers. Further, forensic speech analysis is relevant to i. e. lawsuits where the origin of recorded yells for help is decision-making to distinguish between self-defence or homicide.

Current researches emphasise on text-independent scenarios which e. g., verify on randomised pass-phrases in short duration effort, and on analysing duration-variant speech samples which comprise durations of one second up to many minutes. Thereby, speaker characteristics are modelled by statistical patterns where state-of-the-art research systems prefer template-probe to model-based comparisons, since model-based approaches were shown to be less accurate and having too high computational efforts in duration-variant scenarios. In contrast, template-based systems are known to have disadvantages in short-term scenarios. State-of-the-art researches comprise *identity vectors (i-vectors)* which describe the speaker-characteristic offset to an universal background model.

The applicability of i-vectors will be evaluated in this thesis by comparing i-vector system to well-established model-based approaches on an industry short duration scenario. Thereby, the i-vector approach will be shown not only to operate robust and fast, but also augment existing technologies, such that equal error rates below 0.5% can be achieved. Further, a new duration-mismatch compensation technique will be presented that increases the robustness and performance of i-vector systems in duration-variant scenarios. This new method was evaluated within a current international evaluation of the National Institute of Standards and Technology (NIST) which examines state-of-the-art i-vector systems: the NIST baseline system could be significantly outperformed by a 19% relative-gain in terms of minimum detection costs. Furthermore, this thesis provides a speaker verification framework design which is based on the ISO/IEC 19795-1:2006 *Biometric Performance Testing and Reporting — Part 1: Principles and Framework* standard.

ZUSAMMENFASSUNG

Sprecherverifikation wird immer gefragter für biometrische Lösungen kommerzieller, forensischer und staatlicher Belange. Um Datenmissbräuchen vorzubeugen, gewinnen Telefon-basierte Authentifikationsalgorithmen bspw. zur Datenfreigabe auf mobilen Endgeräten oder zur Nutzervalidierung in Call-Centern mehr und mehr an Popularität. Ferner sind forensische Sprecheranalysen für bspw. Rechtsprozesse relevant, in denen die Klärung des Ursprung eines aufgenommenen Hilfeschreies über Notwehr oder Mord entscheiden kann.

Im Fokus aktueller Forschungsfragen stehen hierbei text-unabhängige Szenarien wie bspw. Kurzzeit-Verifikationen mittels randomisierter Passphrasen und Analysen von Sprachaufnahmen, deren Dauern von unter einer Sekunde bis hin zu mehreren Minuten sehr stark variieren kann. Dabei werden Sprechercharakteristiken anhand von stochastischen Merkmalen modelliert, wobei moderne Systeme mehr auf Vergleichen extrahierter Muster als auf Modell-basierten Analysen aufbauen, da Sprechermodelle in sehr variablen Szenarien als zu ungenau und zeitaufwändig gelten, wo hingegen Muster-basierte Verfahren Nachteile bei Kurzzeitszenarien aufweisen. Aktuelle Forschungen basieren hierbei auf *Identitätsvektoren (i-vectoren)*, welche den charakteristischen Unterschied eines Sprechers zu einem universellen, akustischen Modell beschreiben.

Die Anwendbarkeit von i-vectoren wird im Rahmen dieser Arbeit an einem industriellen Kurzzeitszenario mit bekannten Modell-basierten Verfahren verglichen. Dabei wird nicht nur aufgezeigt, dass der i-vector Ansatz sehr schnelle und akkurate Verifikationen ermöglicht. Im Vergleich zu Modell-basierten Verfahren bieten i-vectoren zusätzliche Informationen, sodass Gleichfehlerraten unter 0.5% erreicht werden können. Ferner wird ein Verfahren zur Steigerung der Robustheit und Performanz von i-vectoren in Szenarien mit sehr stark variierenden Aufnahmedauern vorgestellt, das im Kontext einer aktuellen Evaluation des National Instituts of Standards and Technology (NIST) zu i-vectoren positiv validiert werden konnte: im Vergleich zum NIST System wurde die Verifikationsgüte um 19% gesteigert. Weiterhin wird in dieser Thesis ein Sprecherverifikationsframeworkdesign vorgestellt, das auf dem ISO/IEC Standard 19795-1:2006 *Biometric Performance Testing and Reporting — Part 1: Principles and Framework* basiert.

Information is not knowledge.
Knowledge is not wisdom.
Wisdom is not truth.
Truth is not beauty.
Beauty is not love.
Love is not music.
Music is the best.

— Frank Zappa, 1979.

ACKNOWLEDGMENTS

First of all I would like to thank my supervisors, *Christoph Busch*, *Christian Rathgeb*, and *Herbert Reininger* for giving me the opportunity to write this thesis and research with great freedom at *atip GmbH*. Thanks for the guidance, good advices, and the time you had for me.

Then, I want to thank *Klaus Kasper* for coordinating the *Cooperative Bachelor of Computer Science (KOSI)* giving hundreds of students the possibility to fund themselves through studies, and for his support during international research evaluations.

Within the last six years *atip GmbH* gave me great possibilities not only to improve my knowledge and skills as a computer science student and software engineer, but also for placing the trust in me to participate in the 2012 speaker recognition evaluation of the American National Institute of Standards and Technology and for supporting me to spend an exchange semester at Mälardalens högskola in Sweden during my master's studies. Thanks a lot for all this and to all colleagues, I learned a lot during this time.

Very special thanks to *Martin Wagner* and *Anne Schönwandt* who supervised me in the first four years, introduced me to speech and speaker recognition, and gave me a great guidance to all areas from theory, implementation to evaluation, and more important: for talks about music and yoga lessons. I would also like to thank *Fred Englert* for discussions about phonetics, base guitars and Zappa-related music.

Thanks to *Steve Hegenbart* for researching together, the interesting talks, his support, helpful remarks and the discussions besides work. For her support and discussions I want to thank *Hermine Reichau*.

Many thanks to my parents, *Anke* and *Jochen*, and my brother *Georg* for their never-ending support, trust, love, and patience.

Further, I would like to thank *Max Bernhard* and *Matthias Frank* for the great time beyond the studies.

Thanks to *Noemí Oteiza Izcue* for her love and support and all the great time we shared in Sweden, Finland, Norway, Germany, and Spain during the last year.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AUC	Area Under Curve
BIC	Bayesian Information Criterion
DET	Detection Error Tradeoff
EER	Equal Error Rate
FM	false match
FMR	False Match Rate
FNM	false non-match
FNMR	False Non-Match Rate
GMM	Gaussian Mixture Model
HDF5	Hierarchical Data Format V5
HMM	Hidden Markov Model
i-vector	identity-vector
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LLR	Log-Likelihood Ratio

MAP	Maximum a Posteriori Adaptation
MFCC	Mel Frequency Cepstral Coefficient
NBER	Normalised Bayesian Error Rate
NIST	National Institute of Standards and Technology
OOP	Object-oriented Programming
PCA	Principal Component Analysis
QMF	Quality Model Function
ROC	Receiver Operating Characteristic
SRE	Speaker Recognition Evaluation
UBM	Universal Background Model
VAD	Voice Activity Detection

Part I

BIOMETRIC SPEAKER VERIFICATION

1 INTRODUCTION

In past years speaker recognition has been incorporated in governmental, forensic, and industry applications [1] with a wide-spread scope ranging from court-cases [2] over preventing contact center frauds [3] to key security solutions for high-secure financial transactions [4]. Within conventional speaker recognition systems characteristic traits of an individual's voice are extracted in order to compare them against voice reference data (templates) of known identities, i. e. speakers can either be verified or identified.

1.1 MOTIVATION

The *atip GmbH* designs, implements, and hosts voice applications and supplies many related fields, such as the recognition of speakers. During the last years several Bachelor and Master theses were supervised by atip GmbH aiming continuous-time speaker verification, score normalisations, template protection, and speaker recognition algorithms comprising Support Vector Machines and Joint Factor Analysis (JFA) methods. Concurrent to this thesis, Hegenbart [5] and Billeb [6] were evaluating JFA and template protection methods, respectively. Further, collaborative projects with *Center for Advanced Security Research Darmstadt* (CASED) were started for integrating speaker recognition into the *Modular Biometric Authentication Service System* for Android mobile phones, and fusing speaker and gait recognition systems.

In 2012 the atip GmbH participated in the Speaker Recognition Evaluation (SRE) organised by the National Institute of Standards and Technology (NIST). Techniques mostly submitted by NIST SRE'12 participants comprised JFA and identity-vector (i-vector) approaches where i-vectors were motivated in 2011 [7] as a special case of the JFA. Within the last years the i-vector approach became so well-established due to its extremely fast processing and high performance that NIST hosts the current 2013–14 i-vector challenge which runs in parallel to the time frame of this thesis. Hence, the thesis emphasis is put on the research of i-vectors.

1.2 RESEARCH QUESTIONS

This thesis investigates the performance of i-vector systems in short duration scenarios, which are important to e. g., user verifications during contact-center calls, and how issues on duration-variant scenarios that were reported by the speaker recognition community can be compensated, which is relevant to e. g., continuous-time speaker verifications. Thereby, the following questions will be addressed:

1. *Is the performance of the i-vector approach applicable on short duration scenarios?*
   This is measurable in terms of an Equal Error Rate (EER) below 5%, a *FMR100* under 10% (reporting the genuine mismatch rate on an impostor mismatching rate of 1%), a minimum detection cost below 0.554 (average of the primary systems on NIST SRE'12), a score entropy below $\frac{1}{3}$, and a real-time performance that is not 5% larger than the real-time performance of the baseline approaches.

2. *Do i-vector systems deliver new information to approaches that are known to perform well on short duration samples?*
   The information gain can be shown when the entropy of well-performing systems is reduced due to fusions with i-vector systems.

3. i-vectors were shown in the speaker recognition community to perform robust on long duration samples where on short duration samples immense performance break-downs could be observed; leading to a third research question: *are these mismatches compensable e. g., on the score-level domain?*

1.3 CONTRIBUTION OF WORK

By investigating the questions on short but constant durations of utterances, new technologies will be examined and shown to be applicable as well. Then, current fusion and calibration techniques will be used to improve the performance and robustness of well-known speaker modelling techniques by i-vector-based information, such that significant gains will be demonstrated by e. g., a 56% improved EER and a 36% improved detection cost. Therefore, analyses will be performed on a speech corpus containing sequences of German digits. Further, for reproducibility purposes, a speaker verification framework design will be proposed and implemented where quality is ensured by the application of common tools among the speaker recognition community, and an system design that is based on standards of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).

Evaluations regarding to duration-variant scenarios will be performed on the the NIST-available database from the 2013–14 NIST i-vector challenge. Duration-depending i-vector subspace mismatches will be pointed out and effective compensation techniques will be analysed. In this thesis a new i-vector score normalisation method will be presented, evaluated in comparison to the international speaker recognition community, and shown to yield significant gains in performance and robustness compared to the NIST baseline system.

1.4 ORGANISATION OF WORK

The thesis is organised in three parts: biometric speaker verification in general, evaluation of speaker verification in short-duration scenarios and on varying sample durations, and the appendix with bibliography. The chapters of the following two parts are organised as follows:

- Chapter 2 introduces speaker recognition fundamentals, namely biometrics as a field of forensics and pattern recognition. After introducing the design of biometric systems and biometric performance measurements, forensic evidence strength as performance metric is explained as well. Then, general pattern recognition methods in speaker recognition are discussed.

- Chapter 3 explains speaker verification methodologies beginning from speech signal processing to i-vector extraction, scoring and score post-processing methods.

- Chapter 4 emphasises techniques on short-duration speaker verification and on compensation varying sample duration effects.

- Chapter 5 presents the implemented framework design for biometric speaker verification systems.

- Chapter 6 analyses evaluation results of short-duration experiments, and on compensation varying duration effects on the 2013–14 NIST i-vector machine learning challenge.

- Chapter 7 forms conclusions and points out further research topics.

2 FUNDAMENTALS

This chapter shows speaker recognition related work from forensic theory of optimal Bayes decisions, across characteristic-independent biometric standards, general speech processing technologies, to common pattern recognition methodology. Pattern recognisers have been shown to effectively handle speech signal features in terms of automatic speech recognition [8, 9], language recognition [10, 11], and, speaker recognition [12, 13] as well. Thereby, speech is processed to signal feature vectors, such that biometric systems are able to apply pattern recognition methods, which are well-established among computer scientists.

2.1 FORENSIC EVIDENCE AND BAYESIAN ACTIONS

Forensic science transforms judiciary questions into scientific questions, e. g. *how likely did a suspect interact with a certain environment?*, which may become an essential evidence in the endeavour finding the truth, but does not necessarily answer questions like *did a suspect commit a specific interaction or crime?*. Those questions lead into decisions that are province to courts and not to be made by forensic experts and their investigations [14]. Forensic investigations associate contact of two objects by traces, e. g. blood spots to a concrete individual or a subject's contact with a speech recording device. On contact, traits are transferred between both objects, e. g. due to traces or in form of patterns [14].

In terms of speaker recognition, forensic investigations start with the recorded utterance ω as a transient, digital trace or biometric sample, respectively. A speaker recogniser's task is to associate a sample ω with a concrete individual giving the evidence probability. *Evidence presented in a case at law can be regarded as data, and the issue to be decided by the court as a hypothesis under test* [15]. Biometric systems can measure the evidence by the likelihood ratio of the prosecutor hypothesis H_0 to the defence hypothesis H_A, which argue a subject to have the same identity as the claimed identity or not. Either of both likelihood probabilities are calculated against the sample ω. Evidence reporting is done by a recognisers score S having the interpretation of likelihood ratios [16, 17]:

$$S = \frac{P(\omega|H_0)}{P(\omega|H_A)}. \tag{1}$$

2.1.1 *Bayesian decisions*

Court fact finders, e. g. jury or judge, can use the Bayes' theorem multiplying posterior odds of the likelihood ratios with a factor of prior odds $\tilde{\pi}$, which are *determined by the court after considering other evidence* [15, 16, 17]:

$$\frac{P(H_0|\omega)}{P(H_A|\omega)} = \frac{P(\omega|H_0)}{P(\omega|H_A)} \frac{p(H_0)}{p(H_A)} = S\tilde{\pi}, \tag{2}$$

$$\tilde{\pi} = \frac{p(H_0)}{p(H_A)}. \tag{3}$$

Hence, more complex legal questions can be answered by using combinations according to Bayesian theory of posterior odds and likelihood ratios from several evidences.

2.1.2 *Bayesian actions*

Bayesian decisions can be concluded according to the evidence. These decisions result into Bayesian actions a, e. g. *accept* and *reject* H_0 in terms of a two hypotheses test. In that case one of two actions could be determined by e. g. comparing a score S to a certain threshold t.

Fig. 1 illustrates the relationships of the hypotheses, evidence, Bayesian decisions, Bayesian actions, hypothesis-action-classification, evidence measure performance, and evidence strength according to [18, 19, 20] Thereby, evidence reporting strength can be seen as entropy [18] and the hypothesis-action-classification can be interpreted as a cost matrix where genuine and impostor classifications have zero cost and false matches (FMs) and false non-matches (FNMs) can be set with application-depending costs C_{FM}, C_{FNM}, which are a cost-based interpretation of $\tilde{\pi}$.

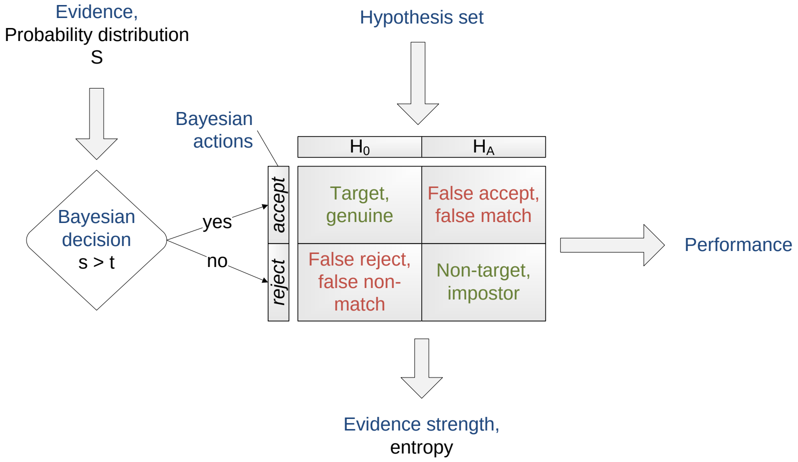


Figure 1: Relationships between Bayesian decision, actions, hypotheses, and their effects according to [19]

2.2 BIOMETRIC SYSTEMS

Biometric systems have the purpose of recognising individuals based on their behavioural and physiological characteristics [21]. This section will give a brief overview of biometric characteristics, a general system architecture from sensory data acquisition to recognition decisions, common error-rates and a brief view on voice-based verification.

Since many communities are dealing with biometrics a harmonised biometric vocabulary [21] has been assembled to reduce vocabulary- and notation-based barriers between all communities. Hence, this thesis is written with respect to a harmonised biometric vocabulary, ISO/IEC 2382-37:2012 [21]: *Information technology — Vocabulary — Part 37: Biometrics*. The complete notation used within this thesis can be found in appendix A, which is also chosen with respect to the speaker recognition community.

2.2.1 *Biometric characteristics*

Biometric characteristics are physiological and behavioural characteristics of individuals by which they can be distinguished, e. g. by *Galton ridge structure, face topography, facial skin texture, hand topography, finger topography, iris structure, vein structure of the hand, ridge structure of the palm, retinal pattern, handwritten signature dynamics* [21].

Tab. 1 classifies selected characteristics according to Jain et al. [1] by universality (U), distinctiveness (D), permanence (P), collectability (C), performance (p), acceptability (A) and circumvention (c), for detailed descriptions, see appendix B. In comparison, physiological characteristics have advantages to behavioural in terms of universality, distinctiveness, permanence, performance, and circumvention. In contrast, behavioural characteristics are easy to collect and higher accepted within the society [1]. By combining systems of different characteristics, sensor types and processing methods as single subsystems to a multi-modal system [21], higher biometric performance might be achieved. Thereby, different modalities imply a difference in characteristic type, e. g. fingerprint and voice. Although, a characteristic type might have physiological and behavioural characteristics.

2.2.2 *Voice-based biometrics*

Voice is referred to as a combination of physiological and behavioural traits, as it is based on physiological characteristics, e. g. vocal tracts, glottal pulse, mouth, nasal cavities, and lips, as well as on behavioural characteristics, such as idiolect, semantics, accent, pronunciation, and prosody [1, 12]. Furthermore, human voice is influenced e. g., by a speaker's medical conditions, emotional states, or ageing, which caus-

Table 1: Comparison of biometric characteristics-based on their technology, excerpt from Jain et al. [1]: high (H), medium (M), and low (L)

Characteristic	U	D	P	C	p	A	c
Physiological							
DNA	**H**	**H**	**H**	L	**H**	L	**L**
Fingerprint	M	**H**	**H**	M	**H**	M	M
Iris	**H**	**H**	**H**	M	**H**	L	**L**
Physiological and behavioural							
Voice	M	L	L	M	L	**H**	H
Behavioural							
Gait	M	L	L	**H**	L	**H**	M
Keystroke	L	L	L	M	L	M	M
Signature	L	L	L	**H**	L	**H**	H

ed Jain et al. [1] to estimate a low permanence. However, Kelly et al. [22] recently proofed that vocal tract ageing effects are easy to compensate[1], thus extracted features describing a characteristic can be assumed to be stable over a long period of time. Jain et al. also assumed a high difficulty to model text- and language-independent speaker recognition systems assuming a 2000 state-of-the-art phone-based speaker modelling. Though, efficient and robust voice modelling approaches were introduced over the last years in terms of JFAs and i-vector developments [7, 23, 24, 25] compensating not only different languages but also sensory effects, such as background noise.

2.2.3 *General biometric system*

Due to the wide-spread variation of biometric characteristics and modalities, many pattern recognition system designs are possible. However, since all have in general same processing flows, a general biometric system framework by components is standardised in ISO/IEC 19795-1 [26]. Fig. 2 illustrates the components and processing flows of a general biometric system according to [26, 27] including data capture, signal processing, data storage to an enrolment database, comparison, and decision making.

2.2.3.1 *Data capturing*

Before either of both recognition procedures, verification or identification, a biometric subject needs to be enrolled. The enrollee presents himself/herself to the system by declaring his/her identity and by sensory interaction to capture the enrollee's characteristics. Thus, a

[1] Ageing compensation is not in the focus of this thesis.

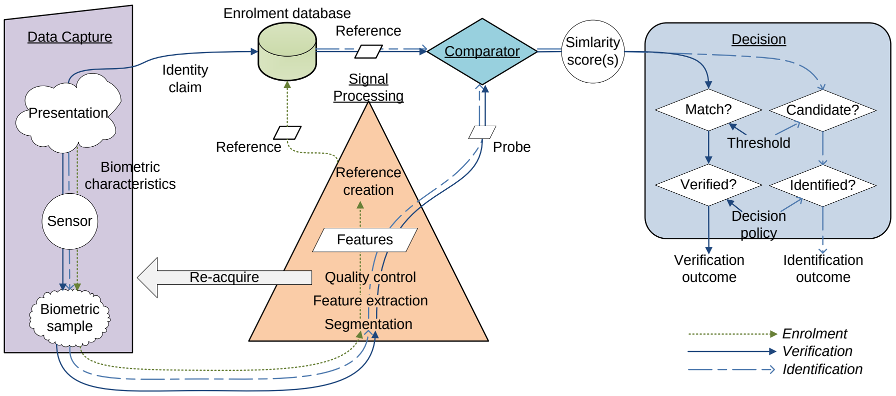


Figure 2: General biometric system, according to [26, 27]

sensor can be a camera capture, e. g. of an ear or a face, or a microphone capturing a humans voice. The captured data is considered as an individual's sample which is afflicted by user and environmental between- and within-variances [1, 13]. Since environmental conditions, e. g. illumination, or background noises, might change, the environment, might change e. g. indoor and outdoor, users change by themselves, e. g. temporary injuries, or growing older, and characteristics of different user should vary, such that individuals can be distinguished.

Data capture of verification and identification differ in terms of an identity claim which is not necessary during identifications, because the identity should be the recognition result, meanwhile a verification outcome determines whether an user's sample matches enrolment templates or models of a claimed identity.

2.2.3.2 *Signal processing*

During signal processing, features are extracted from a biometric sample [26]. Therefore, samples may be segmented with respect to e. g. orientation concerns or specific sub-sequences. ISO [26] defines segmentation as locating a subject's biometric characteristic within a whole sample. Thus, as far as possible de-noised features can be extracted, hence quality control processes are recommended as well in order to re-acquire another sample if necessary.

Depending on the systems mode, enrolment or recognition, feature extractions will produce subject templates or probes, which represent the individual's biometric characteristic. *Sometimes the template comprises just the features* [26]. However, both, templates and models, are referred to as references [21, 26, 27].

References are stored in an enrolment database, which in terms of running applications needs to ensure data security and privacy concerns of e. g. biometric samples and templates [28, 29, 30]. However, this project places emphasis more on research than on full system implementations, hence template protection is not focused on this thesis.

2.2.3.3 *Comparison*

Similarity scores evolve from comparing probes with references, such that likelihoods express their similarity. Comparisons can rely on templates or models [21, 26]: on template-probe comparisons either of both comprise extracted features, hence easy comparison approaches are feasible, such as the cosine distance as similarity measure. On model-probe comparisons, probes are estimated by e. g. statistical models, such that emission probabilities are convenient for likelihood scoring. Each likelihood scoring can be interpreted as evidence in forensic terms, see section 2.1.

For verification only the comparison against the claimed identity's reference is necessary. In contrast, for identification a probe is compared against all references [21, 26].

2.2.3.4 *Decision making*

Decisions are made with respect to the recognition mode: on verification, scores are compared with thresholds to conclude matches or non-matches, which are concerned within decision policies to return verification outcomes of *verified* and *not-verified* [21, 26], likewise *accepted* or *rejected* in terms of Bayesian actions as in section 2.1.2. On identification, identity candidates are determined by threshold-score comparisons. An identification outcome of *identified* or *unidentified* is then concluded by candidate lists with respect to decision policies [21, 26]. Either of both outcomes are Bayesian decisions causing Bayesian actions as mentioned in sections 2.1.1,2.1.2. This thesis places emphasis only on verification.

2.2.4 *Subsystem fusions*

Biometric fusion promises higher accuracy by combining various systems differing in at least one of the following: *sensors, modalities, algorithms, instances or presentations* [31]. According to the technical report ISO/IEC 24722:2007 [31] several simultaneous or sequential presentations can be fused on four levels[2]:

- Sample level,
- Feature level,
- Score level, and,
- Decision level.

[2] The technical report refers also to an additional level: *future undefined fusion method(s)*, which were not concerned within this thesis.

Sample level fusion can be, e. g. sequentially, merging of image series to one image, or simultaneously, capturing characteristics by multiple sensors which are then merged. Fusions on feature level process data from multiple extractors, e. g. histogram gradients and wavelet coefficients. Score level fusions apply score normalisation techniques on scores of multiple comparators, such as linear regression. Decision-based fusions are consolidating logical outcome values of all subsystems by e. g. AND/OR constraints or weighted sums [31]. Fig. 3 shows an exemplary combination of four fusion levels on two samples according to [31].

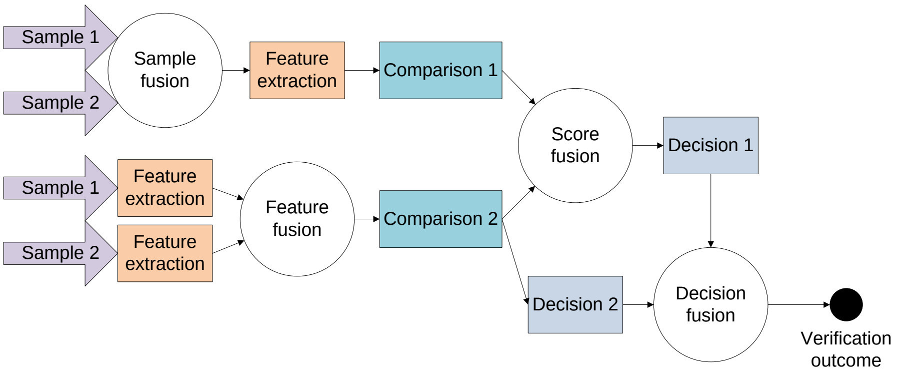


Figure 3: Combination example applying four fusion level, see [31]

Furthermore, fusions are categorised with respect to *multi-modalities* (multiple modalities, algorithms, characteristics, and sensors), *multi-algorithmics*, *multi-instances* (multiple instances of each characteristic), *multi-sensorial*, and *multi-presentation* (same modality, algorithm, characteristic, sensor, but multiple samples e. g. *several frames from a video camera capture of face image* [31]).

However, since all subsystems are applied on statistically dependent data, they should concern correlation *between modalities*, *due to identical samples*, *between feature values*, *among instances due to common operating procedures*, and *among instances due to subject behaviour* [31]. Score level-based multi-algorithmic fusions will be important to this thesis.

2.2.5 *Biometric performance*

Biometric performance addresses recognition accuracy, which is reduced to errors on each processing step from subject presentations to recognition outcomes. Sensors which fail to capture biometric characteristics increase the failure-to-capture (FTC) rate. If features cannot be extracted from captured samples, a re-acquisition is necessary, hence the failure-to-acquire (FTA) rate increases, which is defined by the number of failed feature extraction attempts $N_{\text{failed extractions}}$, the total number of feature extraction attempts $N_{\text{extraction attempts}}$, and the FTC rate [21, 26, 32]:

$$FTA = FTC + (1 - FTC)\frac{N_{\text{failed acquisitions}}}{N_{\text{acquisitions}}}. \tag{4}$$

Uncompleted enrolment cases are measured by the failure-to-enrol (FTE) rate as the *proportion of the population for whom the system fails to complete the enrolment process* [21, 26, 32].

2.2.5.1 *Algorithmic verification error rates*

Biometric verification performance addresses the proportion of false matches (FMs) and false non-matches (FNMs) as mentioned in section 2.1.2. Thereby, the False Match Rate (FMR) is defined by the integral over all impostor scores (H_A is true) greater than a threshold t [1, 21, 26]:

$$FMR(t) = \int_t^{\infty} p\left(S(\omega_{\text{claimant}}|\Omega_{\text{reference}})|H_A\right)dS, \tag{5}$$

with respect to reference and claimant samples $\Omega_{\text{reference}}, \omega_{\text{claimant}}$. The False Non-Match Rate (FNMR) represents the proportion of all non-matched genuines and is defined by the integral over all genuine scores (H_0 is true), such that they were not matched due to a threshold t [1, 21, 26]:

$$FNMR(t) = \int_{-\infty}^{t} p\left(S(\omega_{\text{claimant}}|\Omega_{\text{reference}})|H_0\right)dS. \tag{6}$$

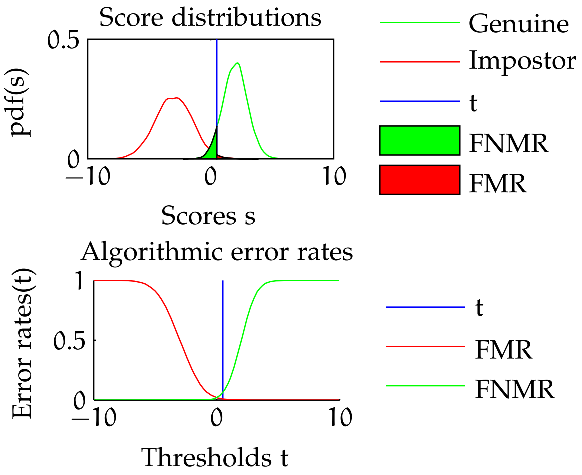


Figure 4: Genuine and impostor score distributions with FNMR, FMR

Fig. 4 shows the relationship between genuine and impostor distributions as probability density functions (pdfs) towards FNMR and FMR on generic scores[3] with respect to a threshold t.

[3] Scores were computed by 10 000 genuine scores having an offset of +2 and 10 000 impostor scores having a scaling of 1.5 with an −3 offset.

Both error rates are representing a biometric system's threshold-dependant tradeoff between security (rejecting/accepting impostors) and user-friendliness (accepting/rejecting genuines). Hence, a high-performance biometric system with security focus should have a very low FMR and a justifiable FNMR [1]. This tradeoff can be represented by e. g. Receiver Operating Characteristic (ROC) curves [33], which compares both error rates threshold-wise, see fig. 5. Thereby, the Area Under Curve (AUC) represents the probability that a biometric system scores a randomly chosen genuine attempt higher than a randomly chosen impostor attempt [33].

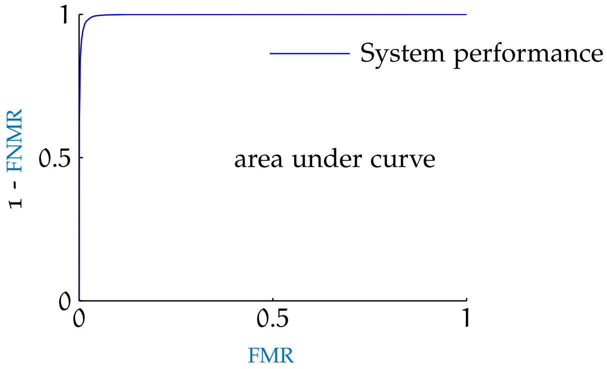


Figure 5: Receiver Operating Characteristic example with area under curve

By displaying both ROC axis in a logarithmic manner and using the FNMR instead of $1 - FNMR$, a biometric system's tradeoff can be displayed in a more natural manner, since all scores are in a logarithmic manner as well [34]. Thus, NIST [34] motivated Detection Error Tradeoff (DET) diagrams such as in fig. 6 according to the example in fig. 4. Due to the logarithmic scale curves appear in a linear manner, hence various system performances can be easy visually compared by their distances. An optimal biometric system would have $FMR = FNMR = 0$. DET plots also provide easily readable performance metrics:

- Equal Error Rate (EER): the rate of FMR and FNMR being equal, thus the EER emphasises neither on secure nor on user-friendly applications.
- *FMR100*: FNMR for $FMR = 1\%$,
- *FMR1000*: FNMR for $FMR = 0.1\%$,

where *FMR100* and *FMR1000* strongly emphasise on secure applications. However, either of both ROC and DET curves have also steppy parts due to a lack of scores. Thus, the statistical significance of error rates needs to be concerned as well. The *rule of 3* [36] *addresses the question "What is the lowest error rate* p *that can be statistically established with a given number* N *of independent identically distributed comparisons?"* [26]. It is defined for a 95% confidence level on N individuals by $p \approx \frac{3}{N}$.

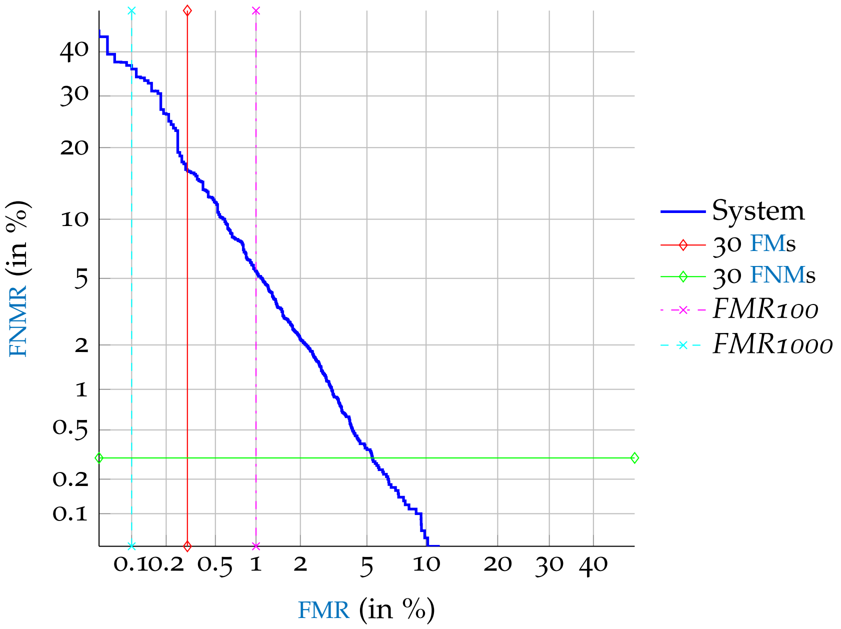


Figure 6: Detection error tradeoff diagram example indicating boundaries by Doddington's rule of 30 [35]

Further, Doddington et al. [35] suggested a *rule of 30* which *states that, to be 90% confident that the true error rate is within ±30% of the observed error rate, there should be at least 30 errors*[4] [26]. Fig. 6 also illustrates the error rate boundaries due to Doddington's rule of 30 and its impact on the separation of insignificant error rate regions. In this case, the *FMR1000* metric is not applicable due to rule of 30 boundaries declaring corresponding FNMR measurements at $FMR < 0.3\%$ as statistically insufficient.

2.2.5.2 *System level verification error metrics*

System level error rates comprise acquisition processes as well [26], hence the False Accept Rate (FAR) is defined by the FMR rate of acquired impostor samples [26]:

$$FAR = FMR(1 - FTA), \tag{7}$$

and the False Reject Rate (FRR) is defined by the FNMR of acquired genuine samples plus not-acquired genuine samples, which where falsely rejected due to failed acquisitions [26]:

$$FRR = FTA + FNMR(1 - FTA). \tag{8}$$

[4] The rule of 30 was introduced within the speaker recognition community [35] and included into ISO/IEC 19795-1:2006 [26].

Furthermore, both metrics can be generalised (GFAR,GFRR) by taking the FTE into account [26]:

$$GFAR = FMR(1 - FTA)(1 - FTE), \tag{9}$$
$$GFRR = FTE + (1 - FTE)FTA + (1 - FTE)(1 - FTA)FNMR. \tag{10}$$

2.3 ROBUST BAYESIAN DECISIONS: ENTROPY OF BAYES ACTS

Besides scores as evidence and systems biometric performances, the evidence strength needs to be reported as well [14, 17]. Furthermore, by interpreting fig. 1 as a cost-matrix, strong evidence pattern recognisers must be assumed to be useful *to make cost-effective decisions in the face of uncertainty* [18, p. 13]. Each Bayes action a, e. g. *accept* or *reject*, is afflicted with entropy [18, 37], because a soft decision score is transferred into a hard decision. Hence, Bayesian actions need to be robust in terms of entropy.

Shannon entropy as a measure for uncertainty of a probability distribution P and its probability mass function p is in general defined for all random variables of $\mathcal{X}$ by

$$\mathcal{H}(P) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = E_P\{-\log p(\mathcal{X})\}. \tag{11}$$

Evidence strength can be seen as the maximum entropy possible of an evidence measure system. The generalised maximum entropy $\mathcal{H}(P^*)$ can be defined as the lowest upper bound of each entropy $\mathcal{H}(P)$, representing all greatest lower bounds of the expected value from all probability mass functions q defined over $\mathcal{X}$:

$$\begin{aligned}\mathcal{H}(P^*) &= E_{P*}\{-\log p^*(\mathcal{X})\} = \sup_{P \in \Gamma} E_P\{-\log p^*(\mathcal{X})\},\\ &= \sup_{P \in \Gamma} \mathcal{H}(P),\\ &= \sup_{P \in \Gamma} \inf_{q \in \Gamma_{\mathcal{X}}} E_P\{-\log q(\mathcal{X})\},\\ &= \inf_{q \in \Gamma_{\mathcal{X}}} \sup_{P \in \Gamma} E_P\{-\log q(\mathcal{X})\},\end{aligned} \tag{12}$$

where Γ is the distribution class of all P, further, $q = p^*$ holds for the maximum entropy $P = P^*$. For robust Bayesian actions, the representative distribution P^* should minimise the worst-case expected logarithmic score (log loss) in terms of evidence strength [37]. Further, the Bayes loss should be measured according to Shannon entropy as well by a logarithmic score, see eq. 11 [37].

2.3.1 *Bayes loss and proper scoring rules*

The Bayes loss $\mathcal{H}_{\text{loss}}(P) \in [-\infty, \infty]$ can be defined by the greatest lower bound of all losses from all Bayesian acts $\mathcal{A}_P$ by a loss function $L(x, a)$ [37]:

$$\begin{aligned} &\sup_{P \in \Gamma} E_P\{-\log q(\mathcal{X})\} = E_P\{L(\mathcal{X}, a)\}, \\ &\qquad\qquad \mathcal{H}_{\text{loss}}(P) = \inf_{a \in \mathcal{A}} E_P\{L(\mathcal{X}, a)\}. \end{aligned} \tag{13}$$

Bayes loss measures the probabilistic prediction accuracy and thus, it can also be seen as Bayes cost, hence loss functions $L_\varphi(x, a)$ are cost functions of the outcome [18, 37]. Loss-based scoring rules can be defined in generic terms by random variables of $\mathcal{X}$ and robust Bayes acts a_Q or in soft decision terms as predictions P and an expected outcome under Q (e. g. the true hypothesis) [37], such that

$$L(x, a_Q) \Leftrightarrow L(P, Q), \tag{14}$$

where $L(P, Q) \leqslant L(Q, Q)$ holds in terms of a similarity metric [38]. Thus, Bayesian loss functions are also proper scoring rules in terms of two-decision problems [18, 38].

Proper scoring rules are defined as special cost functions [18, 38], which were indirectly motivated by Brier [39] to evaluate the goodness of weather forecasts by the observed weather[5]. On strictly proper scoring rules $L(P, Q) \leqslant L(Q, Q)$ holds with equality if, and only if, $P = Q$ [38]. A (strictly) proper scoring rule for logarithmic scores is the log loss $L_{\log}(p, q)$ [18, 37, 38]:

$$L_{\log}(p, q) = -\log p_q \tag{15}$$

which is overall Bayesian action just the Shannon entropy, see eq. 13 [37].

Since the use of Log-Likelihood Ratios (LLRs) is common for likelihood ratio computations, especially in terms of speaker recognition [12, 13, 18], evidence measures will refer to LLRs, thus the logarithmic scoring function will be referred to as the applied proper scoring rule for Bayes loss and Bayes cost.

2.3.2 *Empirical Bayes risk as Bayes error-rate*

The empirical Bayes risk $\mathcal{H}_{\text{risk}}$ in general is based on a Bayes loss function L_φ and cost priors $\Pi = \{\pi_a | a \in A\}$ for each of all Bayesian

[5] Popular proper scoring rule examples are shown in appendix C

actions[6] $\mathcal{A}$, thus the empirical logarithmic Bayes risk $\mathcal{H}_{\text{log-risk}}$ as a special case can be set by $\varphi = \log$, such that:

$$\mathcal{H}_{\text{risk}}(P, \Pi|Q) = \sum_{a \in \mathcal{A}} \frac{\pi_a}{|A|} \sum_{i}^{|A|} L_{\varphi}(p_{a,i}, q_{a,i}), \tag{16}$$

$$\mathcal{H}_{\text{log-risk}}(P, \Pi|Q) = \sum_{a \in \mathcal{A}} \frac{\pi_a}{|A|} \sum_{i}^{|A|} L_{\log}(p_{a,i}, q_{a,i}), \tag{17}$$

where a loss only occurs on wrong decisions, such that the losses of correctly matched (genuines) and correctly non-matched (impostors) are zero. Therefore, the Bayes loss is more defined with respect to genuine and impostor LLRs S and a threshold t rather then to Bayesian actions [18, p. 96], which result due to the comparison $S > t$:

$$L_{\log}(S, t|H_0) = \log(1 + \frac{1}{\exp(S-t)}), \tag{18}$$

$$L_{\log}(S, t|H_A) = \log(1 + \exp(S-t)), \tag{19}$$

where $L_{\log}(S, t|H_0)$:

- vanishes on a good recogniser output, $S_{H_0} \gg t$, due to: $\frac{1}{\exp(S-t)} \approx 0$,
- results in $\log(2)$ on a random recogniser output $S \approx t$, and,
- grows linearly and unboundedly for bad recogniser outputs $S_{H_0} \ll t$,

vice versa on $L_{\text{log}}(S, t|H_A)$. Fig. 7 illustrates the symmetric relation between both loss functions, $L_{\text{log}}(S, t|H_0)$ and $L_{\text{log}}(S, t|H_A)$, which is threshold-independent.

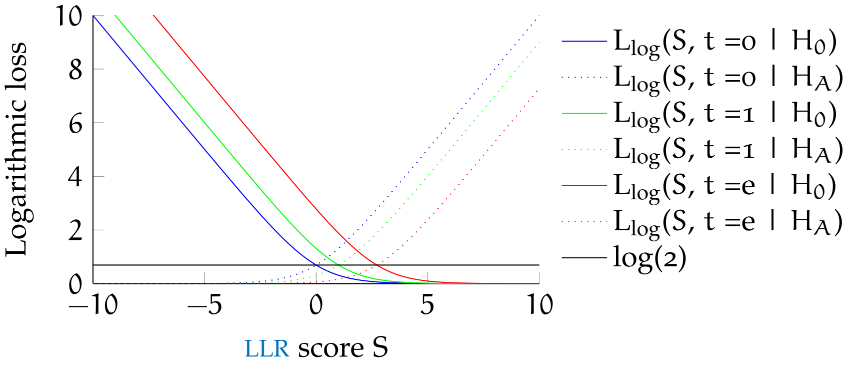


Figure 7: Logarithmic loss of LLRs, for $t = 0, 1, e$

[6] Where $A =$ *accept*, *reject* in binary decision terms, while in terms of assigning patterns to M identities $|A| = M$.

Furthermore, the average of each logarithmic loss can be denoted with respect to verification error functions FNMR and FMR, see [18, p. 75]:

$$\frac{1}{N_{H_0}}\sum_{g}^{N_{H_0}} L_{\log}(S_g, t|H_0) = \mathit{FNMR}(t|\mathbf{S}), \tag{20}$$

$$\frac{1}{N_{H_A}}\sum_{i}^{N_{H_A}} L_{\log}(S_i, t|H_A) = \mathit{FMR}(t|\mathbf{S}). \tag{21}$$

Thus, the empirical logarithmic Bayes risk from eq. 17 can be redefined as the (empirical) Bayes error-rate $\mathcal{H}_{\text{err}}$ on an evidence measure or likewise recognition system $\mathfrak{E}$ [18, p. 55-77]:

$$\mathcal{H}_{\text{err}}(\mathfrak{E}, \pi_{FNM}, \pi_{FM}, t|Q) = \pi_{FNM}\mathit{FNMR}(t) + \pi_{FM}\mathit{FMR}(t), \tag{22}$$

with priors π_{FNM}, π_{FM} for a FNM and FM, respectively. The empirical Bayes error-rate $\mathcal{H}_{err}$ can be interpreted as both: a strict proper scoring rule and generalised cross-entropy [40].

2.3.3 *Forensic prior odds and Bayes thresholds*

In terms of application-specific constraints, such as forensic prior odds $p(H_0), p(H_A)$, both priors π_{FNM}, π_{FM} can be denoted with respect to a synthetic prior π, which models the genuine π and impostor $1-\pi$ probabilities of an application, and error costs C_{FNM}, C_{FM}. Thus, an application domain can be modelled more precisely and its Bayesian error-rate is calculated as [40]:

$$\begin{aligned}\pi_{FNM} &= \pi\, C_{FNM},\\ \pi_{FM} &= (1-\pi)C_{FM},\\ \mathcal{H}_{\text{err}}(\mathfrak{E}, \pi, C_{FNM}, C_{FM}, t|Q) &= \pi\, C_{FNM}\, \mathit{FNMR}(t)\\ &\quad + (1-\pi)C_{FM}\, \mathit{FMR}(t).\end{aligned} \tag{23}$$

Hence, application-dependant priors influence the Bayesian error-rate. However, the Bayes error increases on inappropriate setting of thresholds t which may cause wrong Bayes acts. Thus, Bayes decision thresholds η are motivated by the Bayes theorem in eq. 2 and the prior odds in eq. 3 which substitute t by emphasising more their Bayesian and entropy based motivation. Assuming well-calibrated systems with low entropy an optimal Bayes decision threshold can be constrained on unity of all expected costs if both decisions (Bayes acts) are equally likely, having a default cost of 1 [18, p. 44]:

$$\eta\, \pi\, C_{FNM} = (1-\eta)(1-\pi)C_{FM} = 1 \quad | \,\exists\, 0 < \eta < 1, \tag{24}$$

$$\log\left(\frac{\eta}{1-\eta}\right) = \log\left(\frac{C_{FM}}{C_{FNM}}\frac{1-\pi}{\pi}\right) \quad \Bigg|\log\left(\frac{x}{1-x}\right) = \operatorname{logit}(x),$$

$$\operatorname{logit}(\eta) = \log\left(\frac{C_{FM}}{C_{FNM}}\right) - \operatorname{logit}(\pi). \tag{25}$$

Note that all LLR scores are in logarithmic form, hence η is already applied in logit form, meanwhile the parameters C_{FM}, C_{FNM}, π are not, thus eq. 25 is interpreted as the Bayesian LLR threshold:

$$\eta = \log\left(\frac{C_{FM}}{C_{FNM}}\right) - \text{logit}(\pi). \tag{26}$$

According to eq. 3 only one prior application parameter $\tilde{\pi}$ is necessary instead of π, C_{FNM}, C_{FM}. Such an effective prior $\tilde{\pi}$ can be defined with respect to eq. 23 by [40]:

$$\tilde{\pi} = \frac{\pi\ C_{FNM}}{\pi\ C_{FNM} + (1-\pi)C_{FM}}, \tag{27}$$

$$\eta = -\,\text{logit}(\tilde{\pi}) = -\,\text{logit}\left(\frac{p(H_0)}{p(H_A)}\right), \tag{28}$$

thus eq. 23 can be rewritten as:

$$\mathcal{H}_{\text{err}}(\mathfrak{E}, \tilde{\pi}|Q) = \tilde{\pi}\,FNMR(-\,\text{logit}\,\tilde{\pi}) + (1-\tilde{\pi})\,FMR(-\,\text{logit}\,\tilde{\pi}). \tag{29}$$

Where the effective prior models an application's genuine probability and it's additive inverse $\frac{(1-\pi)C_{FM}}{\pi\ C_{FNM}+(1-\pi)C_{FM}}$ models an application's impostor probability as well as it results in the application-dependant Bayes threshold η.

2.3.4 *Application-dependant entropy*

A systems actual accuracy and Bayes error-rate also depend on the application-dependant threshold η. However, systems having a bad Bayes error-rate $\mathcal{H}_{\text{err}}$ can be highly accurate due to badly distributed LLRs in terms of η: e. g. genuine and impostor scores distributed close to a threshold η might have the same biometric performance as scores that are in a similar manner widely spread, but scores being close around a hard-decision threshold will cause more entropy than widely spread ones.

These score distributions could easily be linearly transformed into better distributed scores according to the thresholds by preserving a systems performance. Such linear transformations are referred to as system calibration [18, 41, 42] which assume a minimum Bayes error rate $\mathcal{H}_{\text{err}}^{\text{min}}$ at a certain threshold $t_{\text{uncal}} \neq \eta$ and a calibration loss resulting in additional entropy $\mathcal{H}_{\text{err}}^{\text{cal}}$ [18, p. 33-37,61,70] — a total error rate $\mathcal{H}_{\text{err}}^{\text{tot}}$ is then defined by the sum of minimum cross-entropy and the calibration information loss:

$$\mathcal{H}_{\text{err}}^{\text{tot}} = \mathcal{H}_{\text{err}}^{\text{min}} + \mathcal{H}_{\text{err}}^{\text{cal}}. \tag{30}$$

An intuitive plotting of calibration loss among several application-dependant Bayes thresholds can only be done after normalisation of application-dependant default error rate effects. For a default system

$\mathfrak{E}_0$ the default error rate changes by different application priors $\tilde{\pi}$. Where $\mathfrak{E}_0$ emits LLRs of $S = 0$, $\mathcal{H}_{\text{err}}(\mathfrak{E}_0, \tilde{\pi}|Q)$ denotes the error of $\mathfrak{E}_0$ and a normalised error $\mathcal{H}_{\text{norm}}$ for $\mathfrak{E}$ can be denoted by [18, p. 76-78]:

$$\mathcal{H}_{\text{norm}}(\mathfrak{E}, \tilde{\pi}|Q) = \frac{\mathcal{H}_{\text{err}}(\mathfrak{E}, \tilde{\pi}|Q)}{\mathcal{H}_{\text{err}}(\mathfrak{E}_0, \tilde{\pi}|Q)} = \frac{\mathcal{H}_{\text{err}}(\mathfrak{E}, \tilde{\pi}, Q)}{\min(\tilde{\pi}, 1-\tilde{\pi})}, \tag{31}$$

which constructs a reference error value of 1 as default entropy by dividing the weighting effects of $\tilde{\pi}$ in eq. 29. Thus, error rates of $\mathcal{H}_{\text{norm}}(\mathfrak{E}, \tilde{\pi}|Q) = 1$ denote a total information loss and a full overall detection cost, which can be interpreted as e. g. costs of wrong Bayesian actions, necessary usage of additional development data, the need of additional subsystems, or an metric to be minimised towards zero [18, 40, 41, 42]. Further, application-dependant entropy values can be compared application-independently [18, 40]. A NBER plot is

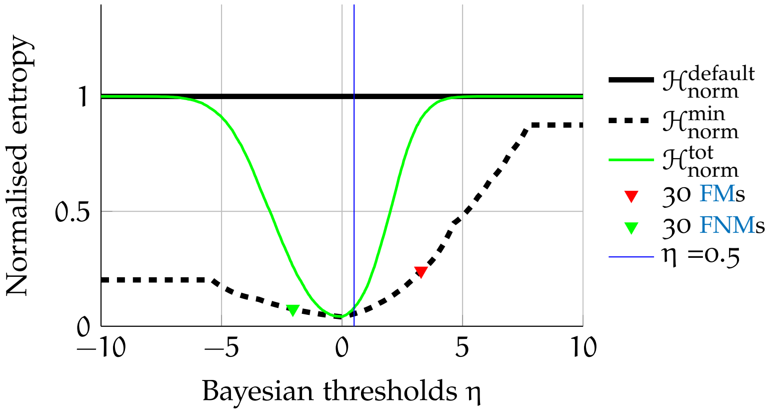


Figure 8: Normalised Bayesian Error Rate (NBER) plot example using the *BOSARIS* toolkit [40]

shown in fig. 8, where the calibration losses from $\mathfrak{E}_{\text{err}}^{\text{min}}$ to $\mathfrak{E}_{\text{err}}^{\text{tot}}$ are presented application-independently and for an example threshold $\eta = 0.5$, with an effective application prior of $\tilde{\pi} = \frac{1}{1+\sqrt{e}} \approx 0.37754$. Furthermore, rule of 30 boundaries can also be plotted with respect to their threshold of occurrence.

2.3.5 *Goodness of log-likelihood ratios*

Application-dependant Bayes error rates are restricted to one effective prior, providing information about how good systems are calibrated for one application, but not application-independently: *Forensic experts should give the court an evaluation, which illustrates the performance of the system, its discrimination value and its robustness to mismatched recording conditions* [16].

The application-independent Bayes error rate of systems can be obtained by integrating out the effective prior, thus a systems cost of LLR scores C_{llr} can be reported as the application-independent goodness of LLRs [18, 40, 41, 42]:

$$C_{llr}(\mathfrak{E}|Q) = k\int_{-\infty}^{\infty} \mathcal{H}_{\text{err}}(\mathfrak{E},\tilde{\pi}|Q)\partial\tilde{\pi},$$
$$= \frac{1}{2N_{H_0}}\sum_{g\in H_0} \mathrm{ld}(1+\frac{1}{e^{S_g}}) + \frac{1}{2N_{H_A}}\sum_{i\in H_A} \mathrm{ld}(1+e^{S_i}). \tag{32}$$

C_{llr} can be interpreted as the scalar expression of the area under NBER curves, as e. g. in fig. 8: $\mathcal{H}^{\text{default}}_{\text{norm}}, \mathcal{H}^{\text{min}}_{\text{norm}}, \mathcal{H}^{\text{tot}}_{\text{norm}}$ where fig. 8 excerpts the interval of relevant operating points[7] $\eta = [-10, 10]$ that correspond to similarity probabilities within $[\approx\ 0.00005, \approx\ 0.99995]$.

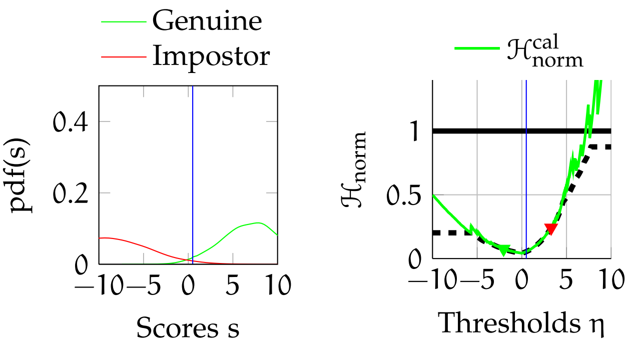


Figure 9: Score calibration example using C_{llr}: score distributions and Bayes error rate

The metric can also be used for linear regression to calibrate scores, hence scores of unknown verification attempts are expected to cause less entropy. Fig. 9 illustrates C_{llr}-based score calibration effects on the scores from fig. 4 as well as on according NBERs[8] [18, 19, 40, 41, 42, 43]: the score distribution is linearly transformed with respect to a certain threshold and according Bayes loss functions, see fig. 7. Thus, only NBERs are effected and biometric performance metrics remain unchanged.

2.4 PATTERN RECOGNITION

The previous sections described how to measure a pattern recognisers performance and entropy, and how pattern recognisers are used in biometric systems. In speaker recognition literature the most popular techniques are based on Hidden Markov Models (HMMs) and

[7] Though, significant operating points are bounded by the rule of 30, such that more accurate C_{llr} computations should emphasise more the according partial-area-under-curve which can be denoted e. g., by 30 FMs and 30 FNMs. However, for the purpose of community comparability C_{llr} computations by the *BOSARIS Toolkit* [40] are emphasised in this thesis.

[8] Please note: the presented curves are trained by and applied on the same scores, hence an unrealistic best case scenario is shown.

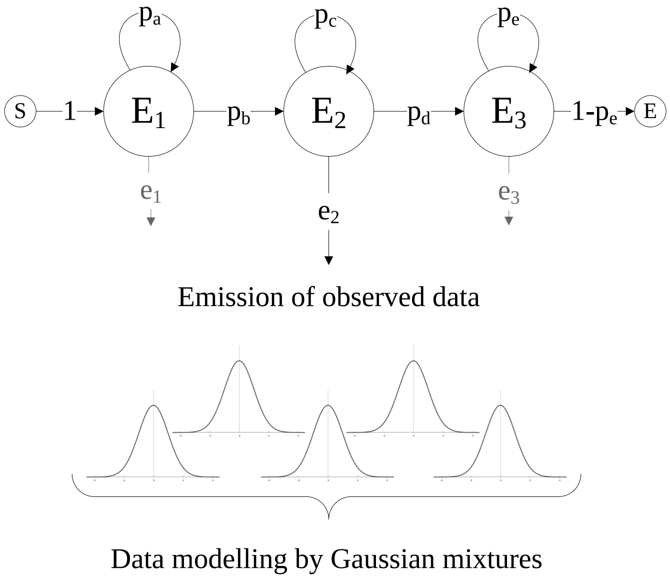
Gaussian Mixture Models (GMMs) to estimate likelihoods for either of both hypotheses H_0, H_A [12, 13]. HMMs were motivated by previous related work on *automatic speech recognition* to model acoustic features with respect to class units of speech sounds (phones) [8, 9]. About the year 2000 GMMs were introduced to speaker recognition for clustering the whole acoustical feature space of spoken languages [44].

Furthermore, biometric pattern recognition need to deal with between- and within-individual variances and thus, with appropriate normalisation approaches [12, 13]. Most recent speaker recognition approaches are applying whitening transformation and Fisher's linear discriminant analysis [13, 45, 46].

2.4.1 *Hidden Markov Models*

A HMM is a two-staged statistical process for discrete-time data, that is based upon a Markov chain, the transition from one state to another [8, 9]. Each possible transition is afflicted with an a-priori *transition probability*, if there is no transition between two specific states, the transition probability is zero. HMM states can be non-emitting and emitting, so they emit any observable data by a-priori statistical modelling, e. g. by Gaussian distributions.

Thus, for certain data observations one state is more likely to emit the observed data than another, hence all possible Markov chains of a HMM can be assigned with an a-posteriori probability according to a certain data sequence by their transition and emission probabilities. The selection of the most likely Markov chain is considered as the second stage of a HMM. Thereby, the actual Markov chain remains *hidden* and only the HMM score is returned [8, 9, 47].



pa
pc
pe
S
1
E1
pb
E2
pd
E3
1-pe
E
e1
e2
e3
Emission of observed data
Data modelling by Gaussian mixtures

Figure 10: Hidden Markov Model (HMM) example

Fig. 10 shows a HMM example[9] with three emitting states, which are modelling data by five Gaussian mixtures with transition probabilities p_{a-e}, emitting states E_{1-3}, emission probabilities e_{1-3}, and non-emitting start and end states S, E. By probabilistic aligning data to HMM states, the amount of possible Markov chains grows quadratically with respect to the amount of states and linearly with respect to the amount of observed data (time series) [8, 9, 47]. Such an alignment can be performed by the Viterbi decoding [8, 9, 47], which is exemplary shown in fig. 11 on a very general speech data presentation: discrete-time data is segmented into sequence frames of same duration, and emitting states are aligned to each frame. For the first frame only one state is possible to align due to the HMM design in fig. 10 and thereafter each transitionable state can be aligned.

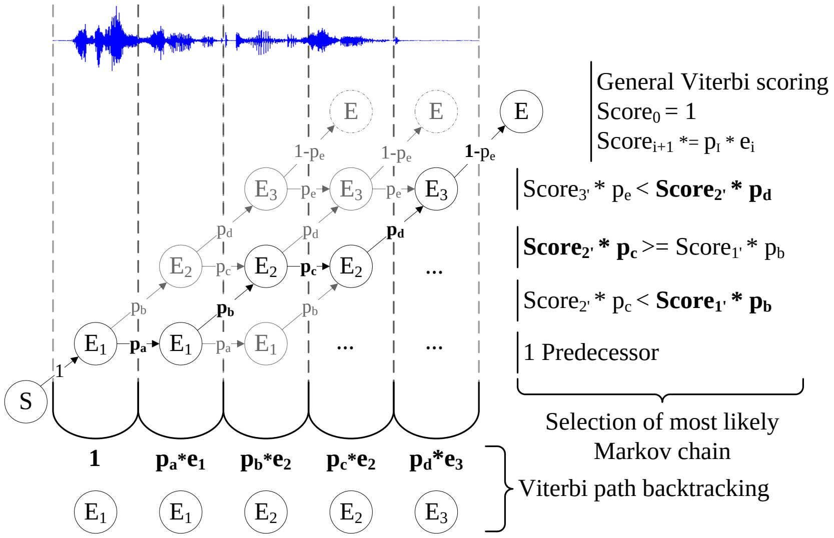

General Viterbi scoring
Score0 = 1
Scorei+1 *= pi * ei
Score3' * pe < Score2' * pd
Score2' * pc >= Score1' * pb
Score2' * pc < Score1' * pb
1 Predecessor
Selection of most likely Markov chain
Viterbi path backtracking
S
E
E1
E2
E3
1
pa
pb
pc
pd
pe
1-pe
1
pa*e1
pb*e2
pc*e2
pd*e3
E1
E1
E2
E2
E3

Figure 11: Viterbi algorithm example with path backtracking, where $p_{a\text{-}e}$ denote transition probabilities and $e_{1\text{-}3}$ denote emission probabilities from the HMM in fig. 10

A Viterbi scoring is performed in the same manner as a Markov chain likelihood is computed by multiplying up all transition and emission probabilities within the chain. On forward computation all chain likelihoods are computed and the most likely Markov chain is selected, meanwhile transition and emission probabilities can be stored as well. Thereby, each emitting state emits data according to its modelling and thus, some frames are more likely to be emitted by one state than by other states, and hence state transition decisions are dynamically influenced by the emission probabilities [8, 9, 47]. The hidden Markov chain or Viterbi path can be obtained by backtracking

[9] The example HMM represents a left-to-right continuous density HMM, which is commonly used for speech processing.

all state transition decisions as shown in fig. 11 with bold most-likely Markov chain or Viterbi path[10] [8, 9, 47].

2.4.2 *Gaussian Mixture Models*

A Gaussian Mixture Model (GMM) can be interpreted as a HMM with one emitting state [44, 47, 48], thus no alignment is needed. A GMM models data frames by joint Gaussian distributions, with C components as Gaussian distributions, corresponding component weights w_c, component means $\vec{\mu}_c$, and component covariances $\boldsymbol{\Sigma}_c$ [44, 47, 48]. Hence, GMMs are generative models able to emit any data with respect to the GMMs dimensionality and to score how likely given data will be emitted by a GMM. Likelihood scores S for feature vectors $\vec{\psi}$ are computed component-wise by Gaussian distributions [44, 47, 48]:

$$\mathcal{N}_c(\vec{\psi}|\lambda) = \frac{1}{\sqrt{(2\pi)^D\,|\boldsymbol{\Sigma}_c|}} e^{-\frac{1}{2}(\vec{\psi}-\vec{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\vec{\psi}-\vec{\mu}_c)}, \tag{33}$$

with D as the feature vectors and multivariate Gaussian distributions dimension. The component likelihoods $\mathcal{N}_c(\vec{\psi})$ are summed up with respect to their component weights resulting into a GMMs λ score S [44, 47, 48]:

$$\begin{aligned} S(\vec{\psi}|\lambda) &= \sum_{c \in C} w_c \mathcal{N}_c(\vec{\psi}|\lambda), \\ &\text{with} \sum_{c \in C} w_c = 1. \end{aligned} \tag{34}$$

A GMM model can be denoted as $\lambda_{GMM} = \{\vec{\mu}, \boldsymbol{\Sigma}, \vec{w}\}$, with a *super-vector* $\vec{\mu}$ of concatenated component means, an overall-components covariance matrix $\boldsymbol{\Sigma}$, and a vector of each components weight $\vec{w}$.

Initial GMM parameters can be estimated after a cluster analysis, e. g. *k-means* or *tree clustering* [47], of a given feature space with respect to a maximum likelihood of the GMM score S to all features: weights are set with respect to component occupancy probabilities, means and covariances are computed component-wise by statistics of the features occupying each component [47].

2.4.2.1 *Baum-Welch statistics*

Baum-Welch statistics [49] are describing GMM (directional) moments with respect to certain feature vectors based upon its components posterior probabilities $P_c(\vec{\psi}) = w_c \mathcal{N}_c(\vec{\psi}|\lambda_c)$ [23, 49, 50, 51]:

[10] Computational optimisation approaches can also be motivated by fig. 11 by e. g. using pruning methods, such as making the state transition decision during forwarding after $n = 1, 2, \ldots, n$ frames and omitting further likelihood computations since the overall Viterbi score will become unlikely.

- zero order statistics $\vec{N_c}(\vec{\psi}) = P_c(\vec{\psi})$ are representing the a posterior probabilities of a feature vector $\vec{\psi}$ (feature-based weight directions),
- first order statistics $\mathbf{F}_c(\vec{\psi}) = P_c(\vec{\psi})\vec{\psi} = \vec{N_c}(\vec{\psi})\vec{\psi}$ are weighting zero order statistics by the observed data (feature-based mean directions), and,
- second order statistics $\mathbf{S}_c(\vec{\psi}) = P_c(\vec{\psi})\vec{\psi}\vec{\psi}' = \mathbf{F}_c(\vec{\psi})\vec{\psi}'$ represent a covariance matrix of the feature vectors weighted by posterior probabilities (feature-based covariance directions).

Furthermore, first and second order Baum-Welch statistics can be centred by GMM means [23, 52]:

$$\mathbf{F}_{c,centered}(\vec{\psi}) = P_c(\vec{\psi})(\vec{\psi} - \vec{\mu}_c), \tag{35}$$

$$\begin{aligned}\mathbf{S}_{c,centered}(\vec{\psi}) &= P_c(\vec{\psi})(\vec{\psi} - \vec{\mu}_c)(\vec{\psi} - \vec{\mu}_c)' \\ &= \mathbf{F}_{c,centered}(\vec{\psi})(\vec{\psi} - \vec{\mu}_c)'.\end{aligned} \tag{36}$$

2.4.3 *Expectation-maximisation algorithm*

For training an initial parameter set towards an optimal training data likelihood weight, mean, and variance values can be iteratively updated by an expectation-maximisation algorithm. The GMM parameter likelihoods rely on its Baum-Welch statistics, thus weights are re-estimated by the average prior zero order statistics, means by the ratio of prior first order to zero order statistics, and variances by the ratio of prior second order to zero order statistics centred by the updated mean variance:

$$\widehat{w_c} = \frac{\vec{N}_c(\vec{\psi})}{T}, \tag{37}$$

with T averaging $\vec{N}_c$ w.r.t. the amount of feature vectors,

$$\widehat{\vec{\mu}_c} = \frac{\vec{F}_c(\vec{\psi})}{\vec{N}_c(\vec{\psi})}, \tag{38}$$

$$\widehat{\mathbf{\Sigma}_c} = \frac{\mathbf{S}_c(\vec{\psi})}{\vec{N}_c(\vec{\psi})} - \widehat{\vec{\mu}_c}\widehat{\vec{\mu}_c}'. \tag{39}$$

2.4.4 *Maximum a posteriori adaptation*

In order to adapt GMMs towards the (acoustical) feature space of biometric characteristics, Maximum a Posteriori Adaptations (MAPs) can be performed by using Baum-Welch statistical information to update each of the GMM parameters: means, covariances, and, weights. Thereby, new parameters $\widehat{w_c}, \widehat{\vec{\mu}_c}, \widehat{\mathbf{\Sigma}_c}$ are set with respect to their prior values weighted by a population priors $\tau_w, \tau_\mu, \tau_\Sigma$ and the observed directional moments as Baum-Welch statistics weighted by the zero order statistic $\vec{N}_c(\vec{\psi})$ [44, 47, 48]:

$$\widehat{w_c} = \gamma \frac{\tau_w w_{c,prior} + \vec{N}_c(\vec{\psi}) \frac{\vec{N}_c(\vec{\psi})}{T}}{\tau + \vec{N}_c(\vec{\psi})}, \tag{40}$$

$$\text{with } \gamma \text{ s.t. } \sum_{c \in C} \widehat{w_c} = 1, \text{ and,}$$

$$\widehat{\vec{\mu}_c} = \frac{\tau_\mu \vec{\mu}_{c,prior} + \mathbf{F}_c(\vec{\psi})}{\tau_\mu + \vec{N}_c(\vec{\psi})}, \tag{41}$$

$$\widehat{\boldsymbol{\Sigma}_c} = \frac{\tau_\Sigma(\boldsymbol{\Sigma}_{c,prior} + \vec{\mu}_{c,prior}\vec{\mu}'_{c,prior}) + \mathbf{S}_c(\vec{\psi})}{\tau_\Sigma + \vec{N}_c(\vec{\psi})} - \widehat{\vec{\mu}_c}\widehat{\vec{\mu}_c}'. \tag{42}$$

Hence, a GMM's log-likelihood is maximised with respect to the observed data $\vec{\psi}$ in terms of Bayesian probability theory. Fig. 12 illustrates MAP effects on two-dimensional generic data after 5, 10, and, 20 iterations for a 2-component GMM: the initial GMM is trained according to initial data (blue) and is then iteratively adapted to the new data (green). MAP terminates if there are no significant likelihood changes to the prior parameter configuration or $n = 20$ iterations are completed [47].

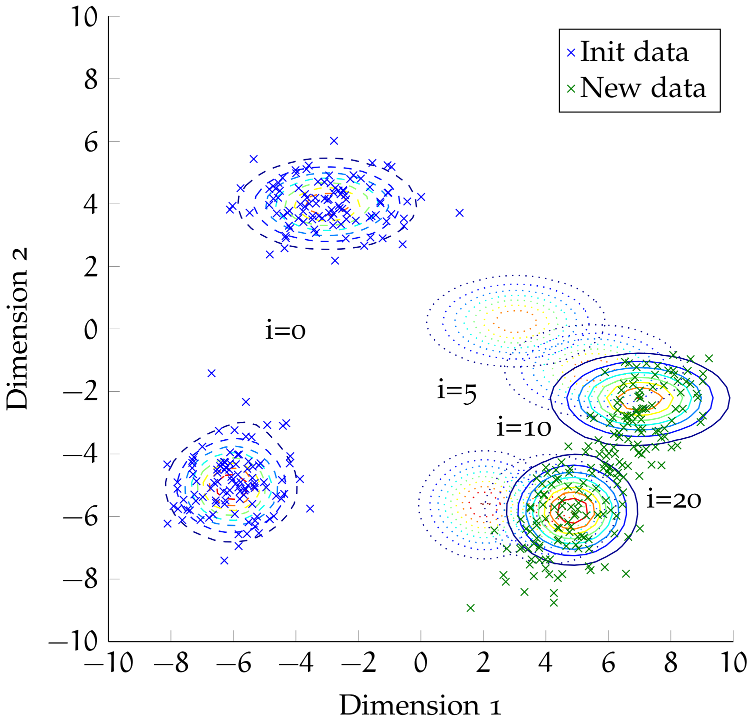


Figure 12: MAP adaptation example with $i = 0, 5, 10, 20$ iterations

2.4.5 *Comparison criteria between Gaussian Mixture Models*

GMMs can be compared using distance and statistical information criteria in quality measurement terms which can be feature data dependent or independent.

As the feature-dependant log-likelihood is used for MAP adaptation, a loss-function can be defined by the negative log-likelihood [53], which can be interpreted as a joint probability distribution function and the joint entropy of a GMM [48, 53]:

$$\mathcal{H}_{GMM}(\lambda|\boldsymbol{\Psi}) = -\frac{1}{|\boldsymbol{\Psi}|}\log S(\lambda|\boldsymbol{\Psi}), \tag{43}$$

which can also be seen as the minimum description length of a GMM [54]. However, a model's minimum description length should also conduct the model's complexity [54], hence the Bayesian Information Criterion (BIC)[11] is motivated, which estimates a GMM's model quality with respect to its entropy and its number of components $|C|$ [54, 55]:

$$\mathrm{BIC}(\lambda|\boldsymbol{\Psi}) = -2\log\left(S(\boldsymbol{\Psi}|\lambda)\right) + |C|\log|\boldsymbol{\Psi}|, \tag{44}$$

hence the BIC can be used to compare two GMMs by their model quality, independent of their kind.

GMM distance metrics commonly rely on the Kullback-Leibler divergence (KL), which is initially defined as the relative entropy between two Gaussian distributions $S(\boldsymbol{\Psi}|\lambda_A), S(\boldsymbol{\Psi}|\lambda_B)$ over features $\boldsymbol{\Psi}$ [56, 57]:

$$\mathrm{KL}(\lambda_A||\lambda_B) = \int S(\boldsymbol{\Psi}|\lambda_A)\log\frac{S(\boldsymbol{\Psi}|\lambda_A)}{S(\boldsymbol{\Psi}|\lambda_B)}\mathrm{d}\boldsymbol{\Psi}, \tag{45}$$

$$\begin{aligned} &= \sum_{c\in C} w_{c,A}\Bigg[\log\frac{w_{c,A}}{w_{c,B}} + \frac{1}{2}\Bigg(\log\frac{|\boldsymbol{\Sigma}_{c,B}|}{|\boldsymbol{\Sigma}_{c,A}|} \\ &\quad + \det(\boldsymbol{\Sigma}_{c,B}^{-1}\boldsymbol{\Sigma}_{c,A}) - D \\ &\quad + (\vec{\mu}_{c,A} - \vec{\mu}_{c,B})'\boldsymbol{\Sigma}_{c,A}^{-1}(\vec{\mu}_{c,A} - \vec{\mu}_{c,B})\Bigg)\Bigg]. \end{aligned} \tag{46}$$

However, the KL is asymmetric. A symmetric KL variant can be constructed by adding $\mathrm{KL}(\lambda_B||\lambda_A)$. Longworth [57, p.52f.] refers to a symmetric KL distance metric based upon the matched-pair upper bound, assuming both GMMs are equally shaped with respect to their components and share weights and variances:

$$\mathrm{KL}^2_{\vec{w}_A=\vec{w}_B,\boldsymbol{\Sigma}_A=\boldsymbol{\Sigma}_B}(\lambda_A||\lambda_B) = \sum_{c\in C} w_c\vec{\mu}'_{c,A}\boldsymbol{\Sigma}_c^{-1}\vec{\mu}_{c,B}, \tag{47}$$

hence GMMs can also be compared feature-independently[12].

[11] The BIC or Schwarz criterion is related to the Akaike Information Criterion (AIC): $2\log\left(S(\boldsymbol{\Psi}|\lambda)\right) + 2|C|$ [55].

[12] Please note: if component-only comparisons are relevant, the Mahalanobis distance $(\vec{\mu}_{c,A} - \vec{\mu}_{c,B})'\boldsymbol{\Sigma}_{c,A}^{-1}(\vec{\mu}_{c,A} - \vec{\mu}_{c,B})$ might be useful as well [57, 58].

2.4.6 *Eigen-analysis*

Eigen-analysis determines non-zero, characteristic vectors, eigen-vectors $\vec{v}$, among a data set matrix $\mathbf{A}$. Thus, *the matrix* $\mathbf{A}$ *stretches the eigen-vector* $\vec{v}$ *by an amount specified by a scalar value, its (characteristic) eigen-value* λ [59]:

$$\mathbf{A}\vec{v} = \lambda\vec{v}. \tag{48}$$

By solving the linear equation $(\mathbf{A} - \lambda\mathfrak{I})\vec{v} = 0$, where $\mathfrak{I}$ denotes the identity matrix, linearly independent eigen-vectors may be obtained, such that λ^{-1} exists, thus data sets can be represented by an uncorrelated linear system:

$$\mathbf{A} = \vec{v}\lambda\vec{v}^{-1}, \tag{49}$$

which is known as *eigen-value decomposition*.

2.4.7 *Whitening*

Whitening is a technique to transform data into more easily usable shapes, which e. g. may allow matched-pair alike data treatments: correlated, differently spaced, and non-unit variant data is transformed into uncorrelated data having zero means and an unit variance of 1 [53].

After means are subtracted from a development data set, its covariance matrix $\boldsymbol{\Sigma}$ eigen-values $\vec{\lambda}$ and eigen-vectors $\mathbf{v}$ are decomposed, hence the variance of the data can be decorrelated and a diagonalised covariance matrix can be written as:

$$\vec{\lambda} = \mathbf{v}'\boldsymbol{\Sigma}\mathbf{v}. \tag{50}$$

The data will be *whitened* by making the uncorrelated variance $\vec{\lambda}$ uniform [60]:

$$\mathfrak{I} = \vec{\lambda}^{-\frac{1}{2}}\vec{\lambda}\vec{\lambda}^{-\frac{1}{2}}, \tag{51}$$

where eq. 50 can be substituted:

$$\mathfrak{I} = \vec{\lambda}^{-\frac{1}{2}}\mathbf{v}'\boldsymbol{\Sigma}\mathbf{v}\vec{\lambda}^{-\frac{1}{2}}. \tag{52}$$

A whitening transformation matrix $\mathbf{W}$ can be created by:

$$\mathbf{W} = \vec{\lambda}^{-\frac{1}{2}}\mathbf{v}'. \tag{53}$$

Thus, a data set $\mathbf{A}$ can be whitened by $\mathbf{AW}$. Please note: $\mathbf{W}$ has the same shape as the original covariance matrix, is semi-positive definite, and symmetric.

2.4.8 *Fisher's linear discriminant analysis*

Fisher's Linear Discriminant Analysis (LDA) is applied to perform dimension reduction by preserving discriminant information and reducing between-variabilities [7]. In contrast to Principal Component Analysis (PCA), LDA assumes the discriminant information to be more within the data mean $\vec{\mu}$ values rather than in its variance [61].

In order to determine discriminant elements $x \in X$ for a class $c \in C$, between and within scatter matrices S_B, S_W are computed with respect to distances between class means μ_c and the overall element means $\bar{x}$ or the class elements x_i, respectively:

$$S_B = \sum_{c \in C} (\mu_c - \bar{x})(\mu_c - \bar{x})', \tag{54}$$

$$S_W = \sum_{c \in C} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)'. \tag{55}$$

By setting between scatters as more important as within scatters, an objective function rating the discrimination accuracy can be denoted for input vectors $\vec{\varphi}$ as [61]:

$$J(\vec{\varphi}) = \frac{\vec{\varphi}' S_B \vec{\varphi}}{\vec{\varphi}' S_W \vec{\varphi}'}, \tag{56}$$

which needs to be maximised for biometric systems. However, by denoting the denominator as $\vec{\varphi}' S_W \vec{\varphi} = 1$, a corresponding Lagrangian optimisation problem L_φ can be defined [61]:

$$\min_{w} - \vec{\varphi}' S_B \vec{\varphi} \text{ s.t. } \vec{\varphi}' S_W \vec{\varphi} = 1, \tag{57}$$

$$L_\varphi = -\vec{\varphi}' S_B \vec{\varphi} + \vec{\lambda}(\vec{\varphi}' S_W \vec{\varphi} - 1). \tag{58}$$

Thus, the following conditions need to hold[13][61]:

$$\begin{aligned}
S_B \vec{\varphi} &= \vec{\lambda} S_W \vec{\varphi}, \\
\Rightarrow S_W^{-1} S_B \vec{\varphi} &= \vec{\lambda} \vec{\varphi}, \\
\Rightarrow S_B^{\frac{1}{2}} S_W^{-1} S_B^{\frac{1}{2}} S_B^{\frac{1}{2}} \vec{\varphi} &= \vec{\lambda} S_B^{\frac{1}{2}} \vec{\varphi}, \\
\Rightarrow S_B^{\frac{1}{2}} S_W^{-1} S_B^{\frac{1}{2}} \mathbf{v} &= \vec{\lambda} \mathbf{v},
\end{aligned} \tag{59}$$

where $S_B^{\frac{1}{2}} \vec{\varphi} = \mathbf{v}$ is substituted.

By decomposing the eigen-problem from eq. 59 a solution $\vec{\lambda}_\varphi, \mathbf{v}_\varphi$ can be found, where on multi-dimensional solutions the most discriminant elements are ranked accordingly to the descend order of $\vec{\lambda}_\varphi$ values. A convenient LDA mapping matrix $\mathbf{M}$ can be created by merging the solution eigen-vectors $\vec{v}_\varphi$ with respect to the corresponding λ_φ order and the reduction dimension D [62]:

$$\mathbf{M} = \mathbf{v}_\varphi(\text{indexes}_{\vec{\lambda}}(\vec{\lambda}_{\varphi,desc}(1, \ldots, D))). \tag{60}$$

[13] According to the first order solution (Karush-Kuhn-Tucker) conditions satisfying L_φ.

2.5 SUMMARY

This chapter comprised forensic evidence, the design of biometric systems, biometric performance and Bayesian entropy evaluations, application-dependant Bayesian thresholds, Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), the Bayesian Information Criterion (BIC), whitening transformation, and Fisher's LDA.

Biometric and entropy performance evaluations rely on falsely matched genuine and imposter attempts. Tab. 2 compares the evaluation metrics on their evaluation scope (score performance, decision entropy, entropy cost of LLR scores), if they can be reported by scalar values or DET/NBER plots, and if they are sensitive towards score calibrations. Furthermore, metric values of the used score example are presented as well.

Table 2: Evaluation metrics overview, according to Brümmer [18, p. 88]

Metric	Scope of Evaluation	Scalar value	Report kind	Calibration-sensitive	Example $\eta = 0.5$
AUC	score	yes	scalar	no	0.9977
EER	score	yes	DET	no	0.0208
FMR100	score	yes	DET	no	0.0539
FMR1000	score	yes	DET	no	0.3447
$\mathcal{H}^{\text{tot}}_{\text{norm}}$	decision	no	NBER	yes	0.0807
$\mathcal{H}^{\text{cal}}_{\text{norm}}$	decision	no	NBER	yes	0.0545
$\mathcal{H}^{\text{min}}_{\text{norm}}$	score	no	NBER	no	0.0545
C^{tot}_{llr}	LLR	yes	scalar	yes	0.2129
C^{cal}_{llr}	LLR	yes	scalar	yes	0.0818
C^{min}_{llr}	score	yes	scalar	no	0.0789

Biometric metrics (AUC, EER, *FMR100*, *FMR1000*) are evaluating score performances, where AUC seems to provide less system comparison information than e. g. *FMR100* or *FMR1000*, which represent the user-friendliness on highly-secure performances rather than comparing an overall probability if its more likely to reject a genuine or accept an impostor. The EER delivers an upper-bound for either low FMRs or low FNMRs.

Entropy-based metrics indicate whether systems need more additional information for making forensic-reliable decisions. Since forensic evidence is influenced by additional prior knowledge, such as e. g. a genuine attempt probability, effective priors $\tilde{\pi}$ and according Bayesian thresholds η are effecting the application-dependant metrics: $\mathcal{H}^{\text{tot}}_{\text{norm}}, \mathcal{H}^{\text{cal}}_{\text{norm}}, \mathcal{H}^{\text{min}}_{\text{norm}}$. Where differences between $\mathcal{H}^{\text{tot}}_{\text{norm}}$ and $\mathcal{H}^{\text{min}}_{\text{norm}}$ indicate badly calibrated scores, hence score calibration techniques might be applied to biometric systems to reduce their decision entropy resulting in $\mathcal{H}_{\text{norm}}^{\text{cal}}$, an optimisation of $\mathcal{H}_{norm}^{tot}$. Thus, the relative gain from $\mathcal{H}_{\text{norm}}^{\text{cal}}$ to $\mathcal{H}_{\text{norm}}^{\text{min}}$ can provide an insight to how much a systems entropy can be reduced to a minimum entropy given by proper scoring rules based on a weighted sum[14] of the algorithmic error rates FMR and FNMR.

Calibrations may rely on proper scoring rules such as C_{llr}, which is an application-independent entropy metric evaluating the goodness of LLR scores. According to each of the application-dependant metrics $\mathcal{H}_{norm}^{tot}, \mathcal{H}_{norm}^{cal}, \mathcal{H}_{\text{norm}}^{\text{min}}$ application-independent metrics can be derived: $C_{llr}^{\text{tot}}, C_{llr}^{\text{cal}}, C_{llr}^{\text{min}}$. However, LLR decision-based metrics are calibration-sensitive: they can be optimised or weakened, depending on the data shift between development and evaluation data, but if the relative gain between C_{llr}^{cal} and C_{llr}^{min} is low, calibration can be denoted to be adequate. C_{llr} values can be reported either as scalar values or as the area under $\mathcal{H}_{\text{norm}}$ curves, hence all Bayesian error-rates can be reported by NBER plots, but for comparability reasons C_{llr} will be reported by scalar values. This thesis will refer to results on EER, *FMR100*, $\mathcal{H}_{\text{norm}}^{\text{min}}$, and C_{llr}^{tot} using DET and NBER plots.

[14] Please note: this weighted sum has application-dependant weights.

3 SPEAKER VERIFICATION

This chapter provides an overview on current speaker verification methodologies. Speaker verification applies pattern matching algorithms on voice-based characteristics [13]. Characteristic information about the voice of a subject can be obtained by extracting speech signal features. Thereby, the acoustical features depend on the speaker and the spoken text. Acoustical features can be clustered by e. g. the smallest speech units delivering information, the *phones* [12].

Different speakers have different subspaces within an universal acoustical feature space [7, 12, 23]. The subspace offset from the universal cluster describes the direction vector of a sample which depends on the verbalised text of a speaker. In order to obtain relevant speaker-only features, these vectors are analysed towards characteristic factors [7, 23]. The factor-analysed features are referred to as identity-vectors (i-vectors) [7].

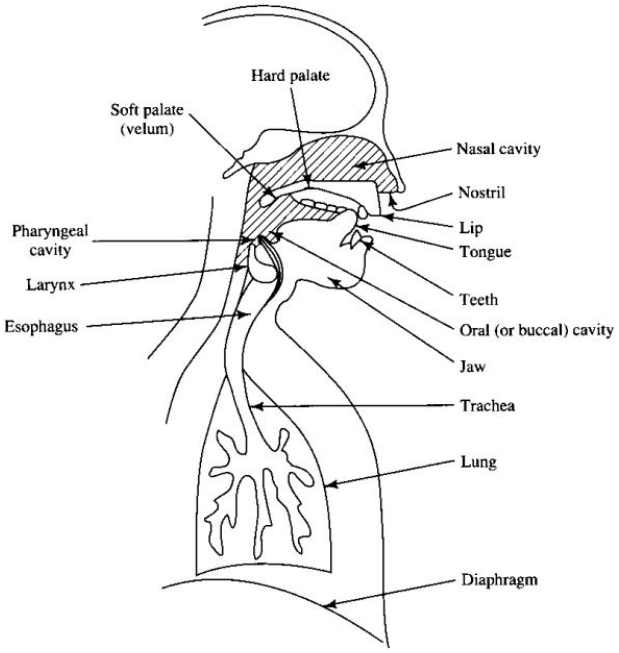
In biometric terms i-vectors are features and in machine learning terms they are patterns. Either way, an i-vector comparison can be simply performed by measuring the angle between reference and probe i-vectors [7]. More advanced comparators also take speaker variabilities into account [24, 63], e. g. due to changes in environmental noise and recording locations, different microphones as capturing devices (between-speaker variabilities), varying microphone distances, emotional and physical states, or ageing (within-speaker variabilities).

Further, system scores are normalised to augment the recognition results with a-priori knowledge of similar verification attempts [24, 64]. Verification systems are calibrated in order to reduce recognition entropy, and several verification systems can be fused into one recognition system accumulating all advantages of each of the subsystem recognition accuracies [18, 40].

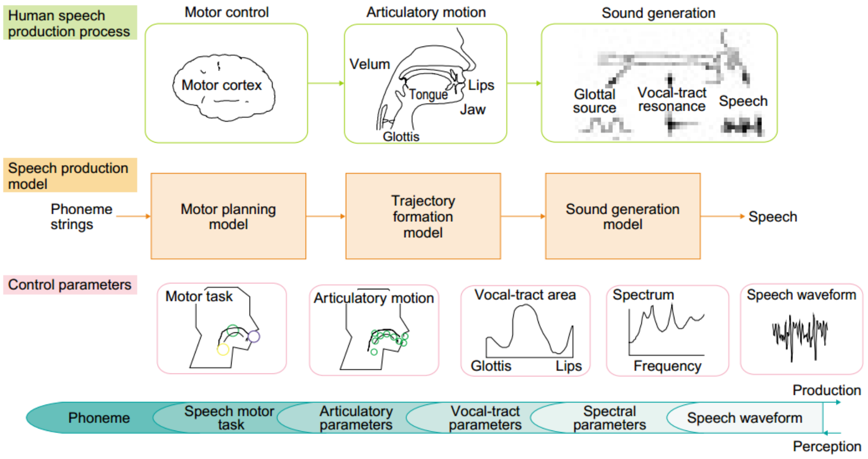
This chapter is organised as follows: in sections 3.1 and 3.2 speech processing and a clustering of the acoustical space are introduced, then previous speaker modelling approaches are explained in section 3.3 in an abstract way. Sections 3.3 and 3.4 present factor analysis methods on speech patterns, the extraction of i-vectors, and i-vector scoring approaches. Further, section 3.4 explains score normalisation, calibration, and fusion techniques.

3.1 SPEECH PROCESSING

Human speech is produced by articulatory motions controlled by the motor cortex [65]. These articulatory motions start with pressing air out of the lungs, which passes the glottis where a speaker's voice main characteristic traits are shaped by the vocal folds[1] [66]. Speakers control whether the air moves through the nasal or oral cavity or both by their velum [66]. The nasal cavity's shape influences the shape of nasal phones, while air passing the oral cavity is influenced by the oral cavity's shape, tongue motions, jaw shapes, and the lip's shape and motions [66]. Fig. 13 illustrates the vocal anatomy and the motion control on speech production which, in terms of biometrics, form physiological and behavioural characteristics.



(a) Anatomy of the articulation system, see [67]



(b) Speech production and motor control, see [65]

Figure 13: Articulation anatomy and speech production

[1] Vowels are produced with open vocal tract, consonants with complete or partial closure of it [66].

The production of speech results in shaped air pressure which is due to human perception assigned with language-dependant meanings [65]. By recognising varying air pressure, humans are able to communicate, and distinguish between speakers [14]. Changes in air pressure are used for pattern recognition as well as in automatic systems for speech and speaker recognition [8, 9].

Microphones measure changes in air pressure and the changing velocity by e. g., electricity changes in terms of capacitor voltage as pressure affects the distance between two plates, or carbon granules resistance as higher pressure affects a membrane pushing granules together, such that electric resistance is lowered [68]. Either of both, capacitor and carbon microphones, are well-known to hand-held and cordless telephone industry [68]. Fig. 14 shows recorded membrane changes as raw values over time of a 2012 NIST SRE [69] enrolment sample.

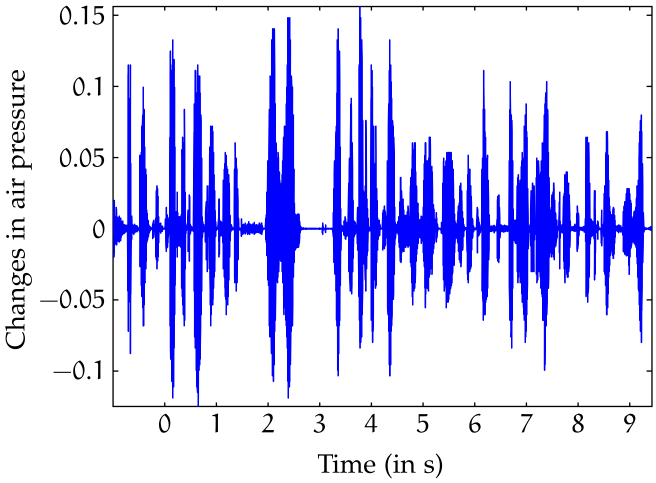


Figure 14: Sample with raw valued speech in waveform

Raw valued speech signals contain every sound that was recorded. Segments containing speech are more important to speaker recognition rather than segments without speech nor with noise [12, 13]. Hence, voice activity needs to be detected.

3.1.1 *Voice activity detection*

Voice Activity Detection (VAD) removes sample segments that are likely not to contain speech. These segments are considered to be those with non-low changes in pressure or likewise the signals energy [12, 13]. This section explains briefly how the VAD algorithms work. The signal energy E of all raw samples ω_{raw} is computed by

$$E(\omega_{\text{raw}}) = \langle \omega_{\text{raw}}(t), \omega_{\text{raw}}(t) \rangle = \int_{-\infty}^{\infty} |\omega_{\text{raw}}(t)|^2 dt, \tag{61}$$

where t represents the time [47]. The segment length can be defined by a (short) duration in milliseconds or likewise by an amount of N_{seg}

raw sample values: the duration of phones can vary from 30 ms to 200 ms [70] and the speech is recorded by a specific sampling rate [47], hence the N_{seg} should be set with respect to the minimum phone aspiration duration and the Nyquist-Shannon sampling theorem[2] in order to not exclude speaker information. Further, two energy-adaptive thresholds[3] are set denoting confident the presence of speech t_{speech}, and the presence of speech pauses with environmental noise t_{noise} [12, 71].

However, VAD needs to be robust against temporal signal impulses and tolerant towards phones having low signal energy. Thus, in order to label the signal as speech and non-speech each threshold needs to be satisfied by a minimum amount of segments: N_{speech} segments having $E > t_{speech}$ and $N_{non\text{-}speech}$ segments having $E \leqslant t_{noise}$ [12, 71]. Fig. 15 illustrates VAD processing on the speech sample from fig. 14.

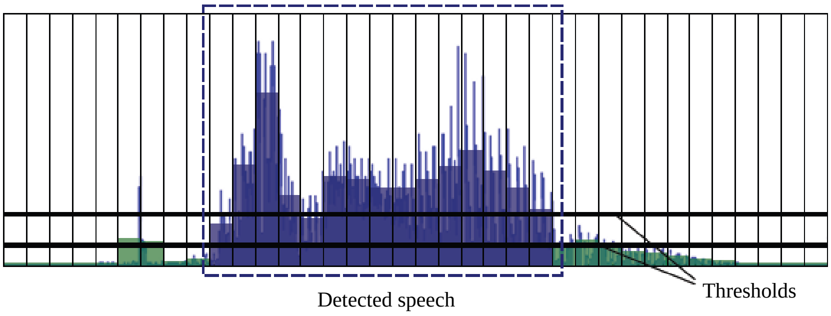


Figure 15: VAD applying segment-wise two adaptive thresholds on signal energy, see Hegenbart [5]

The resulting VAD filtered sample comprises speech data, and has a reduced duration which in computational terms of e. g. phone calls can averagely downsize 40–60% of the sample values, since two sides share the whole sample duration [5]. Hence, further feature extractions process less data with more relevant speaker information.

3.1.2 *Speech signal features*

Speech signal features in speaker recognition are influenced by behavioural and physiological characteristics. Fig. 16 compares feature-domains according to the summary of Kinnunen and Li [12] on ad-

[2] The Nyquist-Shannon theorem is a fundamental theorem in information theory. The maximum bandlimit of a countable sequence of samples is not greater than half the sampling rate. In given terms: $N_{seg} \leqslant \frac{30}{2}\text{ms} = 15\text{ms}$.

[3] The thresholds are adaptive in order to deal with quiet and loud speech and they are adaptively set rather with respect to the overall sample value than to local energy values, because environmental noise might be non-linear and suddenly impulses due to in-/decreasing energy values, e. g. due to coughing, need to be considered as well.

vantages and disadvantages of different speech characteristics from high-level spoken language features to short-term spectral features.

Short-term spectral features comprise more physiological characteristics of phone articulations which have strong related motions between speakers, otherwise spoken language would not work. However, articulated sounds vary in different physiological traits such that individual speakers can be distinguished [12]. By analysing mid- and long-term speech signals, prosodic features can be obtained such as the glottal pulse[4], (phone/word) durations, and rhythm, thus more behavioural traits which still strongly depend on the physiology, e. g. shapes of the glottis muscles. In contrast, high-level features describe behavioural traits that were learned within social and language backgrounds, such as the idiolect, semantics, accent, dialect, and pronunciation [12].

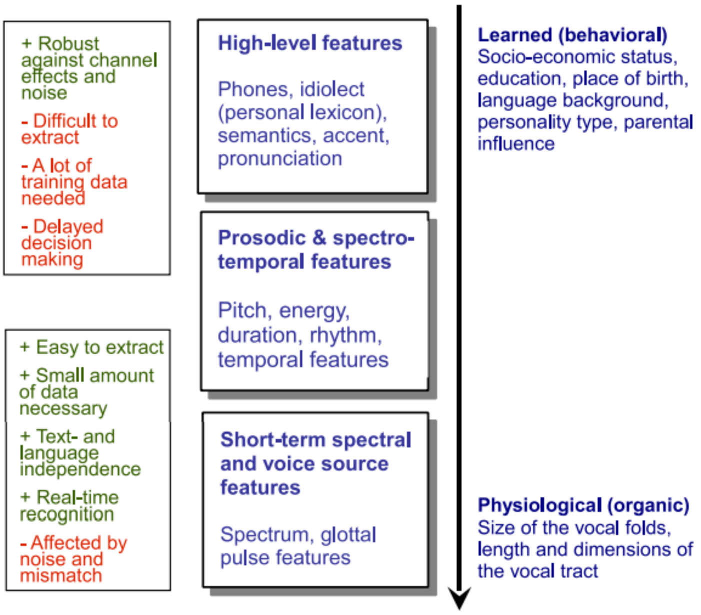

+ Robust against channel effects and noise
- Difficult to extract
- A lot of training data needed
- Delayed decision making
+ Easy to extract
+ Small amount of data necessary
+ Text- and language independence
+ Real-time recognition
- Affected by noise and mismatch
High-level features
Phones, idiolect (personal lexicon), semantics, accent, pronunciation
Prosodic & spectro-temporal features
Pitch, energy, duration, rhythm, temporal features
Short-term spectral and voice source features
Spectrum, glottal pulse features
Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence
Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract

Figure 16: Speech signal feature overview: from high-level behavioural to short-term spectral (physiological) features according to Kinnunen and Li [12]

High-level features are not influenced by noise and channel effects, though they are conceded to require a lot of training data, are difficult to extract under high-computational costs. Since they are based inter alia on idiolect and semantics, they are language-dependant [12]. However, short-term spectral features can be extracted in real-time and are text- and language-independent, since they are emphasising

[4] The glottal pulse is related to the pitch and the fundamental frequency. In phonetic terms the averaged glottal pulse is usually referred to, since it depends on the characteristics of verbalised phones which are strongly influenced by a humans constitutional state such as exhausted, thrilled, and pleased which humans recognise by the same characteristic, the changes in the fundamental frequency of another person [12].

more on phone articulation parts rather than on a complete phone. Though, these features measure signal noise and channel effects as well. Prosodic features have advantages and disadvantages of both, e. g. the rhythm describes behavioural traits, but is affected by signal noise as well, hence raw features can be extracted fast, but need more post-processing that may include human post-editing [12].

According to Rose [72] ideal features in forensic speaker recognition comprise:

- large between-speaker, small within-speaker variability (distinctiveness),
- robustness against signal noise and channel effects,
- frequent occurrences in natural speech (universality),
- easy measurable (collectability),
- difficult impersonation (low circumvention),
- invariances, e. g. health and ageing (permanence).

They are very related to the quality classification of biometric characteristics by Jain et al. [1], see section 2.2.1 and appendix B. Since short-term spectral features are referred to as being easy computable from small data amounts and emphasising on physiological characteristics rather than learnable high-level features [12]. For the purpose of researching secure systems, low circumvention verification systems are aimed, thus features describing the speech signal spectrum are emphasised on in this thesis.

3.1.2.1 *Spectral features*

Spectral features are analysed in short-term segments [12, 47]. Hence, waveforms are given for each segment. The waveforms have huge variations depending on different phones as well as on physiological-caused between-speaker variations. In order to extract spectral features easily, their frequency spectrum is analysed with respect to discriminative frequency bandwidths [12, 13]: humans having more low-frequent changes in the speech signal than high-frequent changes are articulating with less changes in air pressure, and hence with less vocal effort. E. g. women have higher voices than men [13]: the female vocal folds are shorter than the male vocal folds, in order to circulate the same amount of air, they need higher-frequent vocal fold motions, and thus the fundamental frequency of women is higher than the fundamental frequency of men [66]. Further, different speakers can be distinguished by comparing patterns of the whole frequency band: due to the physiological shape of speakers articulation systems and behavioural articulation motions, changes in speech signals like (air) pressure occur with different frequencies on different humans [12, 13].

After transforming time-based segments into frequency domain, frequencies are commonly processed in an auditory-based manner which is motivated by the human ability to distinguish voices by speech perception which begins at the signal transformation of the human ear[5] [12, 13, 73]. In order to obtain features describing a short-term speech segment, the frequency domain features need to be transformed back into the time domain, hence spectral features are obtained by a so called cepstral[6] analysis. Here the spectral features will be further denoted as cepstral features [12, 13]. Fig. 17 gives an overview on cepstral feature extraction through a block scheme: after frequency domain analysis, an auditory-based processing is performed, such that cepstral coefficients can be obtained by cepstral analysis [73].

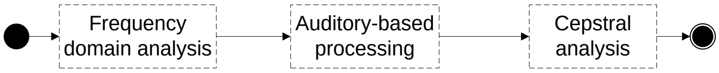


Figure 17: Processing flow for extracting cepstral features, see [73]

3.1.2.2 *Frequency domain analysis*

The transformation of time domain issues into the frequency domain is generally performed by the *Fourier transform* [9, 12, 47]. The Fourier transform decomposes time domain waveforms into overlapping frequencies which can be represented by sine and cosine functions. Thereby, the spectral density of the sine and cosine functions will be obtained. In order to determine the energy spectral density[7] of a frequency ξ within a given continuous-time waveform function $f(t)$, each waveform sample is normalised by the expected amplitude of ξ at a certain time t. The energy spectral density is then determined by integrating the normalised waveforms over time, further sine and cosine dependencies can be substituted using Euler's formula[8]:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} \frac{f(t)}{\cos(2\pi\xi t) + i\sin(2\pi\xi t)} dt = \int_{-\infty}^{\infty} f(t)e^{-2\pi i \xi t} dt \quad (62)$$

where i denotes the imaginary unit. However, since speech signals are in a discrete-time domain having T samples, eq. 62 can be redefined as the discrete Fourier transform

$$\hat{f}(\xi) = \sum_{t=0}^{T-1} f(t)e^{-2\pi i \xi \frac{t}{T}}. \quad (63)$$

[5] Other features such as *perceptual linear prediction coefficients* [12] rely more on articulation-based processing.

[6] *Ceps*tral is an anagram of *spec*tral illustrating the frequency domain transform and the *backwards* transform.

[7] The energy spectral density is also known as the power spectrum.

[8] Euler's formula: $e^{ix} = \cos(x) + i\sin(x)$.

Thus, short-term speech samples can be transformed into frequency domain obtaining information describing the frequency impact on the observed waveform, the energy spectral density [9, 12, 13]. Fig. 18 shows the Fourier transformed sample of the raw sample of fig. 14.

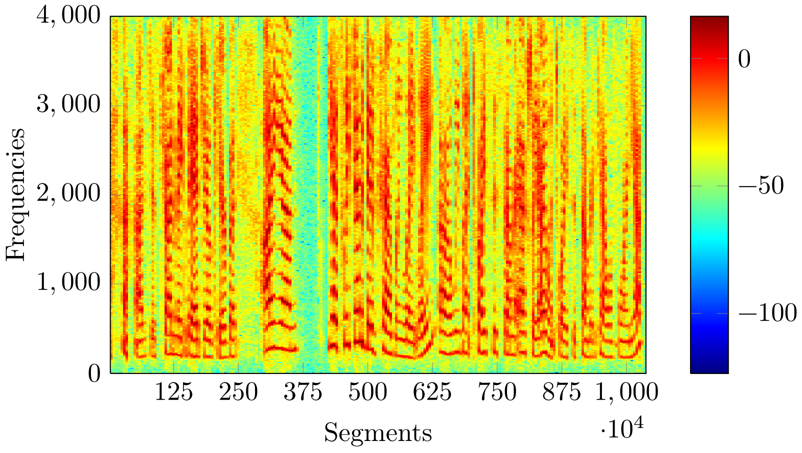


Figure 18: Energy spectral density of the short-term analysed speech sample

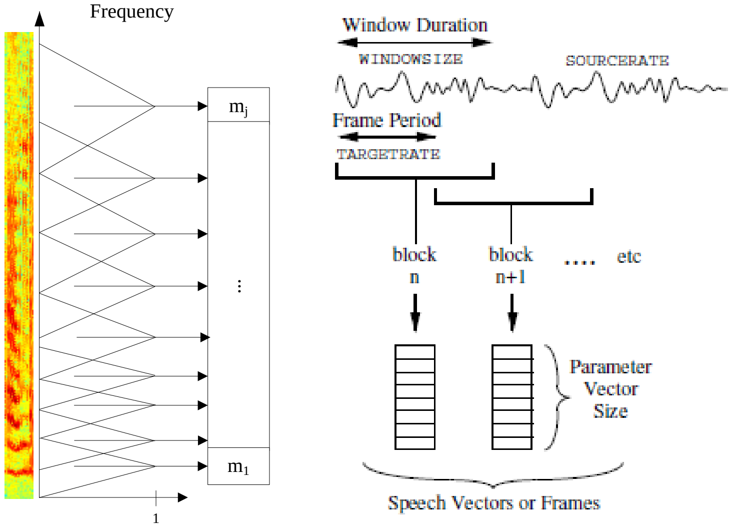
3.1.2.3 *Auditory-based filter-banks*

The spectral density measures the signal as it is, but not as it is recognised by the human ear [9, 47]. During the last century psycho-acoustical studies were performed regarding to the human perception of sound [73]. The perception studies placed emphasis on the frequencies of equal pitch increments resulting in logarithmic dependencies: the *melody-scale* (mel-scale) which is based on pitch comparisons is defined by [9, 47]:

$$m(\xi) = 1127 \log\left(1 + \frac{\xi}{700}\right). \tag{64}$$

By rescaling the frequency ranges, the energy information becomes more discriminative within certain regions. These regions are commonly (in the speaker recognition community) processed by using 20–24 triangular band-pass filters[9] which are equally spaced according to the mel-scale [73], hence they are in a logarithmic manner in the frequency domain as shown in fig. 19a. The filtered energy in each mel-scaled band represents the acoustical human perception of the short-term signal, usually the logarithmic filterbank amplitudes $m_{j \in N_{\text{banks}}}$ are used [9, 12, 13, 47, 73].

[9] Further perception research on critical band rates showed that there are 24 critical bands of hearing [47, 66] which are related to the mel-scale, hence using up to 24 band-pass filters continues the auditory approach.



(a) Mel-scale-based triangular band-pass filter

(b) Feature extraction by sliding segment windows

Figure 19: Extracting MFCC features according to [47]

3.1.2.4 *Cepstral coefficients analysis*

In order to obtain cepstral features, the calculated filterbank amplitudes are transformed into the time domain, an *inverse discrete cosine transform* is performed which is a special case of the inverse of the discrete Fourier transformation from eq. 63, thus only real numbers are kept. Cepstral features c are computed by [9, 47]:

$$c_k = \sqrt{\frac{2}{N_{\text{banks}}}} \sum_{j=1}^{N_{\text{filter-banks}}} m_j \cos\left(\frac{\pi k\,(j - 0.5)}{N_{\text{banks}}}\right), \tag{65}$$

where k denotes the k-th cepstral feature that should be obtained to describe the short-term speech signal in terms of MFCCs, and j iterates over the triangular band-pass filter-banks illustrated in fig. 19a. The speaker recognition community usually refers to extract between 12 and 19 MFCCs which are also supposed to be statistical independent [9, 12, 74, 75, 76]. For preserving as much speech information as possible, cepstral features are computed from overlapping sliding segment windows [47], as shown in fig. 19b.

Further, the MFCCs can be augmented using the logarithmic energy of the short-term segment [12, 47], see eq. 61, and by appending overtime information as MFCC delta (velocity) Δ, delta-delta (acceleration) ΔΔ, and third differential (jerk) ΔΔΔ coefficients which are computed by regression [47]:

$$\Delta\, c_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \tag{66}$$

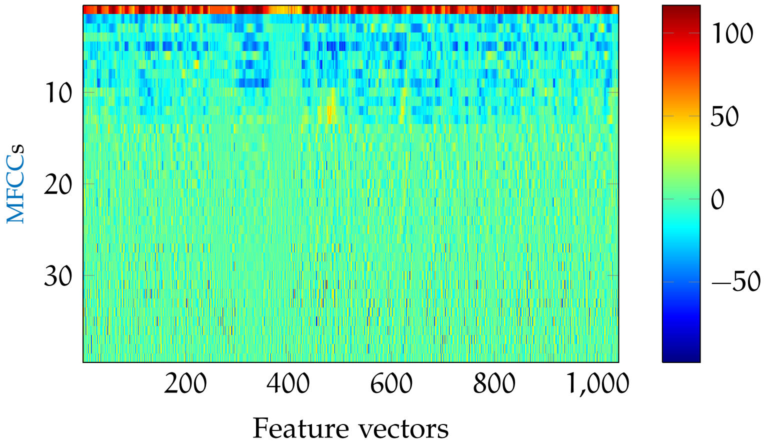


Figure 20: Cepstral features of a sample

where Θ denotes the Δ-window size. In order to reduce over-time effects of continuous non-speech noise, the static non-Δ MFCCs can be zero-normalised, so that linear signal effects vanish [13, 47]. Fig. 20 shows 39 extracted short-term features for the example sample as 12 MFCCs augmented with the signal energy comprising Δ and ΔΔ features.

3.1.3 *Feature normalisation*

Features need to be normalised in order to reduce between-sample or likewise within-speaker variances. An appropriate normalisation for MFCC features was introduced by Pelecanos and Sridharan [77]: MFCC values extracted from mid-term speech samples are Gaussian distributed, hence between-sample mismatch effects can be reduced by warping them into a target distribution according to their equal source probability. Fig. 21 illustrates the *feature warping* idea.

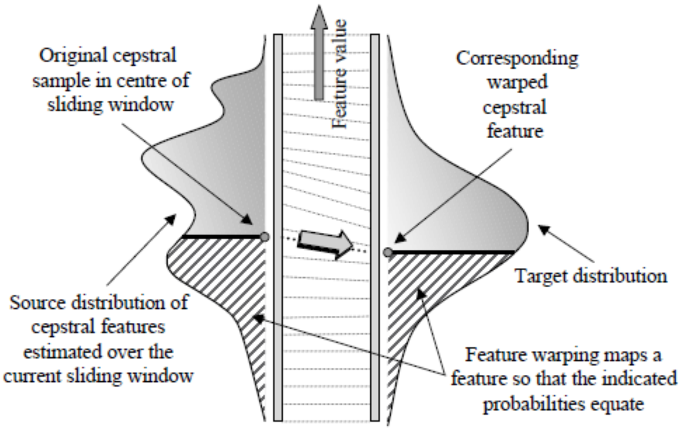

Original cepstral sample in centre of sliding window
Feature value
Corresponding warped cepstral feature
Target distribution
Source distribution of cepstral features estimated over the current sliding window
Feature warping maps a feature so that the indicated probabilities equate

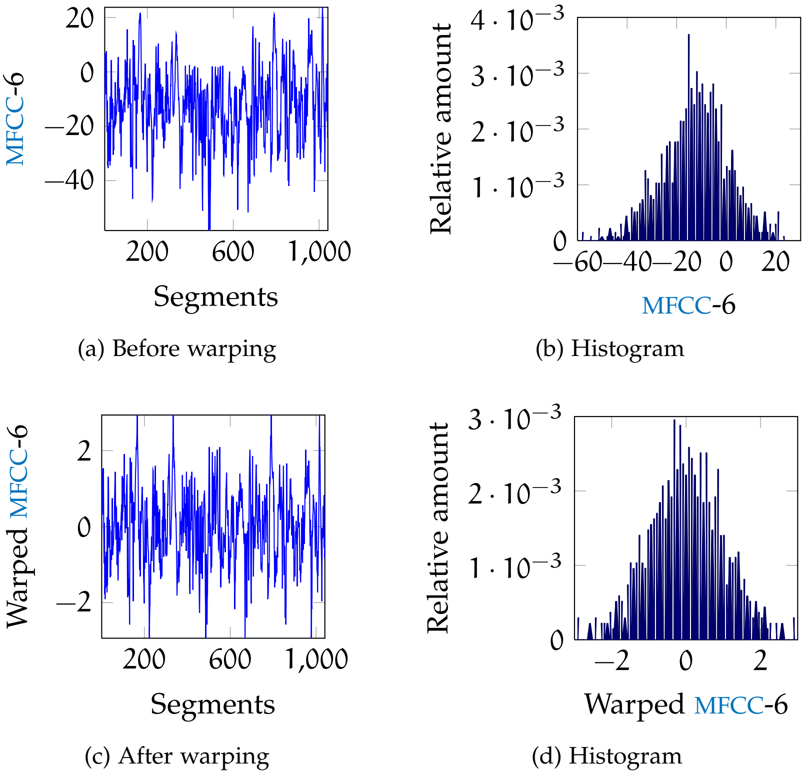
Figure 21: Feature warping according to [77]

Feature warping is commonly applied on 3-seconds sliding windows[10] [75, 77, 78]. Within each feature sequence only one feature value, the window-center feature, is warped with respect to the features around it. Therefore, the feature's rank R within the window size $N_{features}$ is mapped according to the equal-probability of a standard Gaussian as the target distribution where the warped feature value w is the accordingly expected value [77]:

$$\frac{\frac{1}{2} + N_{features} - R}{N_{features}} = \int_{z=-\inf}^{w} \mathcal{N}(z) \partial z. \tag{67}$$

Thus, a mapping can be initially created which relates $N_{features}$ feature ranks to $N_{features}$ warped values.

An example of feature warping is shown in fig. 22: the first 3 seconds of the example samples first and second MFCCs are shown which are warped independently, the according histograms, and the warped features with their according histogram. The unwarped features occur as Gaussian with noise, which is reduced by feature warping [77].



(a) Before warping

(b) Histogram

(c) After warping

(d) Histogram

Figure 22: Feature warping on the example sample: normalised feature value intervals (standard Gaussian)

[10] On 10-20ms feature extraction windows this would accord to 150–300 samples.

3.2 CEPSTRAL SPACE ANALYSIS

MFCC features represent the acoustical or likewise the cepstral space of speech samples which comprise not only speech signals. Hence, the cepstral space is shaped by: environmental and channel noise, and the speech which depends on the phones of the verbalised text, and the speaker's voice, see fig. 23. Where non-speech noise can occur due to linear and non-linear effects.

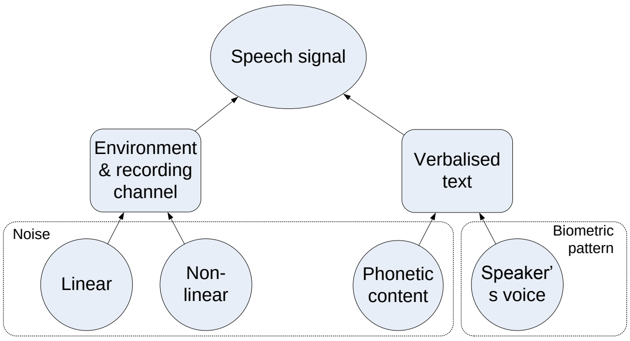


Figure 23: General speech signal influences

Assuming an existing cepstral space cluster, each speech sample is situated in a subspace depending on channel effects, the phonetic content, and articulation characteristics of a speaker [7, 12, 23]. In the following speaker verification will be explained which is based on a speaker-independent acoustical space clustering, such that speaker-dependant offsets can be determined and used for template-probe comparisons.

3.2.1 *Clustering acoustical features*

As explained in section 3.1 MFCC features are Gaussian distributed and are not significantly correlated. Hence, GMMs are appropriate for clustering 39–60 dimensional or likewise multivariate feature vectors representing 12 MFCCs, the signal energy with their Δ and $\Delta\Delta$ values. If a GMM is trained by speech data of many speakers having varying texts (phonetic contents), different languages, various durations, different environmental, and channel effects, then the cepstral background is universally modelled for all speakers, thus such GMMs are referred to as Universal Background Models (UBMs) [12]. Further, these UBMs have diagonal covariance matrices, since MFCC features of one feature vector can be assumed to be mutual independent. An example cluster and a sample's subspace are shown in fig. 24.

However, all UBMs depend on their training data e. g., if an UBM is trained by subjects speaking the same texts (which may vary within a phonetic close-content), this UBM may be trained text-independently, but its application scope is due to its training more restricted to an

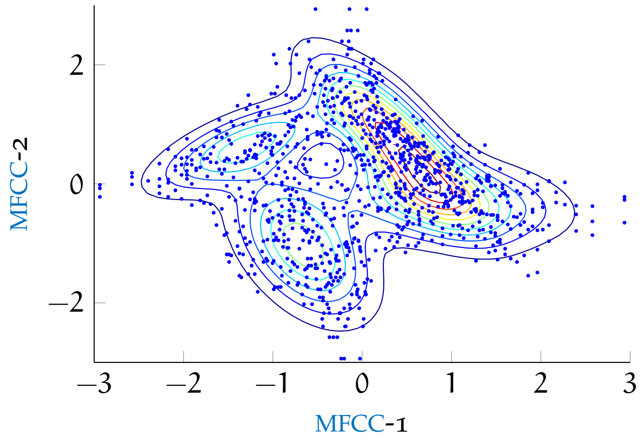


Figure 24: Cluster example with four components

according close-content scenario rather than to text- and language-independent application scenarios, because information about phones cannot be modelled if they are not represented within the UBM development dataset. Also majority feature occurrences of, e. g. phones, channel noises, male or female gender and certain age ranges, will cause UBMs to emphasise on those effects during training and likelihood estimations as described in section 2.4.2, rather than delivering an independent cepstral space cluster which also represents feature vectors application-independently [79]. In contrast, on close-text-content application scenarios as digit pass-phrases, UBMs are supposed to operate better, if they are trained by according close-content development datasets, because modelling phonetic content that never will be requested from UBMs will cause noise and thus increase entropy. Meanwhile close-content UBMs can also suffer from over-represented phonetic content e. g., if biometric users are asked to repeat random pass-phrases which prevents replay attacks on verification processes by including user knowledge[11]. Thus, speech features need to be selected by reducing UBM over-fitting and preserving the features natural distribution.

3.2.2 *Intelligent feature selection*

Hasan and Hansen [79] assumed feature cohorts which have very small Euclidean within-distances e. g., MFCC features representing single phones. Thus, they proposed an intelligent feature selection (IFS) algorithm which is based on the Euclidean distance: only feature vectors are selected for further processing that deliver new information towards the last selected feature vector(s). Therefore, the feature vectors are processed one by one and the current distance between a

[11] This thesis is not emphasising how to fraud biometric speaker recognition systems by e. g., real-time voice modulations, hence research and industry scenarios are emphasised.

feature vector and its prior vector is compared to the distribution of all distances: if the distance is too low to gather new information, it is rejected, otherwise new information can be gained and the feature vector is selected.

Euclidean distances are Chi-square distributed. Hasan and Hansen [79] proposed to recursively update the Chi-square distribution's parameters after each selected or rejected (processed) feature vector. Thereby, feature vectors are selected, if the distance to the last selected feature vector is within the upper α-quantile of the current updated distance distribution. Thus, feature vectors likely to contain new information about the cepstral space are selected by an adaptive distance threshold. In fig. 25 the idea on cepstral cohort groups is illustrated where only one representative from each group needs to be selected, so that over-fitting can be reduced.

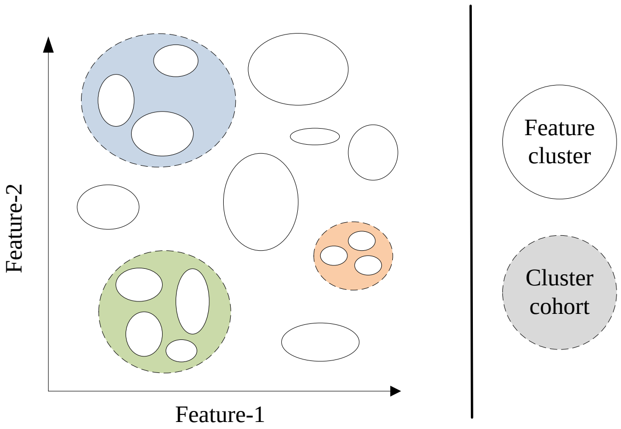


Figure 25: Schematic feature space with cohort groups according to [79]

3.3 SPEAKER SUBSPACE ESTIMATION

Speakers are modelled by their enrolment samples, such that an UBM is adapted using the MAP algorithm described in section 2.4.4 towards a speaker's subspace. This speaker modelling approach is essential for speaker verification techniques of the last 15 years which comprise e. g., HMMs, GMMs, and i-vectors [12, 44, 64].

In this section model advantages and disadvantages of HMM- and GMM-based speaker subspace estimations are discussed. Which comprise short-term content modelling and also full text-independent approaches.

3.3.1 *Short-term content subspaces*

In speech recognition phones and words are modelled using HMMs [8, 9], see section 2.4.1. Thus, in speaker verification terms HMMs are defining short-term content spaces for e. g. phones, words or digits. Hence, a HMM-based text-independent speaker verification system first recognises which sample parts belong to which phones and secondly performs according likelihood score estimations.

During enrolments, each speaker needs to give such speech samples that all HMM models can be sufficiently trained by MAP adaptation. HMMs are usually used for modelling words or phones, hence spoken enrolment texts need to contain sequences such that according speech data can be extracted and used for speaker enrolments. Thereby, HMM UBM weights and variances are not updated, because they comprise much more phone variation information than within-speaker variations. Thus, only the means as speaker characteristics giving information about articulation frequencies are MAP adapted.

A HMM score S_{HMM} is calculated by the LLR between the probability of a probe's features $\boldsymbol{\Psi}$ being emitted from an enrolled speaker model λ_χ and the probability of a probe sample being emitted from the UBM model Λ_{UBM}. Assuming already aligned states $r = 1, \dots, R$ of the HMM model λ, such that all probe features $\boldsymbol{\Psi}$ can be written by according sequences $\boldsymbol{\Psi}_{r=1,\dots,R}$, a LLR score is computed by:

$$S_{\mathrm{HMM}}(\boldsymbol{\Psi}|\chi) = \sum_{r=1}^{R} \sum_{\vec{\psi} \in \boldsymbol{\Psi}_r} \log \frac{S(\vec{\psi}|\lambda_{r,\chi})}{S(\vec{\psi}|\Lambda_{r,\mathrm{UBM}})}. \tag{68}$$

HMMs are able to model co-articulation effects as well, because of their state-based signal modelling kind: each state is adapted to speaker-specific articulation phases by MAP adaptation, see sections 2.4.1 and 2.4.4. Hence, they are able to take effects of short- and mid-term features according to Kinnunen and Li [12]. However, this means HMMs need a lot of training data to estimate sufficient UBM statistics which vary from language to language, hence HMMs have disadvantages on multi-lingual scenarios. Further, HMMs need to know the spoken text or articulated phones, which raises especially in long-term scenarios a lot of non-speaker recognition dependent issues e. g. false phone recognitions. Also the Viterbi alignment complexity[12] in HMM-based speaker verification increases quadratically by the amount of HMM states, and hence the recognised phone or word models do so as well.

In contrast, GMMs are feature-wise evaluating samples without taking spectro-temporal effects as much into account as HMMs[13], thus the Viterbi algorithm needs not to be applied.

[12] HMM-based verification is only a border area to this thesis' work, and thus not explained in full detail. The complexity of the Viterbi algorithm is denoted in terms of all states of all concatenated HMMs S and the amount of frames T which can be seen as the sample time: $O(|S|^2 T)$.

[13] Spectro-temporal effects are only influencing GMM-based speaker recognition in terms of Δ MFCCs on very short spectral sequences. However, HMMs are influenced by spectro-temporal features in terms of e. g., phone- or phrase-depending effects.

3.3.2 *Text-independent speaker subspace*

GMM-based speaker verification techniques create one UBM cluster representing the whole acoustical space text-independently [12]. The effects of phones, channels, and speakers are modelled according to the expectation maximization algorithm explained in section 2.4.3. Hence, impacts of certain speech signal effects can only be roughly assigned to certain GMM components, since various effects affected the UBM parameters in order to obtain huge effect-independence and thus, universality.

In a GMM-UBM system, speakers are enrolled by MAP adapting the UBM by their enrolment samples as described in sections 2.4.4 and 3.3.1. Only mean UBM parameters are MAP adapted, because UBM mean values are averaging speaker articulation frequencies, meanwhile weight and variance parameters are more compensating phonetic, and channel effects which cannot be sufficiently estimated by a few enrolment samples having restricted variational effects [12, 13, 44, 48]. Thus, GMM mean values represent samples and speakers. The mean concatenated vector-form is referred to as *supervector* $\vec{\mu}$ [12, 13].

During verification, a probe sample's features $\boldsymbol{\Psi}$ are scored by their emission probability $S_{\mathrm{GMM}}(\boldsymbol{\Psi}|\lambda_{\chi})$ of the speaker model λ, see eq. 34 which satisfies H_0. Further, the emission probability $S_{\mathrm{GMM}}(\boldsymbol{\Psi}|\Lambda_{\mathrm{UBM}})$ of the UBM model Λ_{UBM} satisfies H_A, which can be written in a feature vector-wise processing as [44, 48]:

$$S_{\mathrm{GMM}}(\omega|\chi) = \sum_{\vec{\psi}\in\boldsymbol{\Psi}_{\omega}} \log \frac{S(\vec{\psi}|\lambda_{\chi})}{S(\vec{\psi}|\Lambda_{\mathrm{UBM}})}. \tag{69}$$

The score $S_{\mathrm{GMM}}(\omega|\chi)$ represents the LLR score as the GMM-UBM comparator's result.

3.3.3 *Supervectors as templates and features*

Supervectors are in combination with an UBM a complete speaker model, hence on GMM-UBM systems they are speaker templates. However, in terms of the UBM being a cluster supervectors denote the subspace's cluster center. The difference between sample and UBM supervectors represent sample offsets from the initial UBM cluster:

$$\textit{UBM-offset} = \vec{\mu}_{\omega} - \vec{\mu}_{\mathrm{UBM}}. \tag{70}$$

i-vector-based speaker verification systems are interpreting supervectors as features that were extracted by UBMs in order to determine the UBM offsets [7]. By extracting speaker-characteristic values from the sample-dependant offsets more accurate template-probe comparisons are expected.

3.3.4 *Characteristic cepstral space offsets*

Characteristic UBM cluster offsets were analysed with respect to HMM supervectors [80], and GMM supervectors [51] where Kenny et al. [23] supposed a PCA-based supervector mapping into an *eigenvoice* space. Eigenvoice models assume discrete channel effects [23], such that for given template and probe supervectors one channel-dependant vector is synthesised by a speaker vector (eigenvoices) and an offset vector (eigenchannels) which is afflicted by Gaussian noise. Hence, random speaker samples can be estimated and classified.

In 2008 Kenny et al. [23] motivated JFA of GMM supervectors in order to treat speaker-characteristic factors and channel-dependant factors separately. Comparing eigenvoices with JFA, the eigenvoices approach models sample variabilities, the JFA approach models speaker variabilities [23]. JFA analyses speaker factors $\vec{y}$, channel factors $\vec{x}$ and residual factors $\vec{z}$ by extracting speaker-dependant eigen-alike vectors. Thereby, eigen matrixes **V**, **U**, **D** , **T** map speaker-dependant supervector variabilities into a lower dimensional space [23, 50, 51]:

$$\vec{\mu}_{\omega} = \vec{\mu}_{\mathrm{UBM}} + \mathbf{V}\,\vec{y}(\chi) + \mathbf{U}\,\vec{x}(\omega) + \mathbf{D}\,\vec{z}(\lambda_{\chi}) \tag{71}$$

where $\vec{\mu}_{\mathrm{UBM}}$ is the speaker-independent UBM supervector, χ,λ_{χ} denote the speaker, and the speaker model, respectively. Hence, the sample-dependant UBM offset is decomposed into speaker and channel factors.

However, Dehak et al. [7] found out that the eigenchannel matrix **U** contains also eigenvoice information, hence they suggested a *total variability* factor analysis in order to reduce redundant information which yielded biometric and computational performance gains.

3.3.5 *Total variability factor analysis*

The total variability approach models a sample's supervector $\vec{\mu}_{\omega}$ on the basis of a speaker-independent supervector $\vec{\mu}_{\mathrm{UBM}}$ which is afflicted by the variation of a speaker-characterising i-vector [7]. Thereby, a total variability matrix **T** which intentionally synthesises all variations that can occur on a speech sample and maps a low-dimensional i-vector $\vec{\imath}$ in the supervector space:

$$\vec{\mu}_{\omega} = \vec{\mu}_{UBM} + \mathbf{T}\,\vec{\imath}(\chi). \tag{72}$$

In terms of a supervector – UBM offset modelling as in eq. 70, analogically eq. 72 can be rewritten as:

$$\vec{\mu}_{\omega} - \vec{\mu}_{UBM} = \mathbf{T}\,\vec{\imath}(\chi). \tag{73}$$

Thus, the speaker-characteristic i-vector is determined by decomposing a supervector's UBM offset with respect to a prior modelled total variability matrix.

The total variability matrix is a-priori computed on a development dataset, e. g. the UBM data [81]. In order to efficiently estimate proper i-vectors according to eq. 73 sample supervectors $\vec{\mu}_\omega$ need to be estimated and centered by the UBM supervector $\vec{\mu}_{UBM}$. Supervectors are concatenated GMM mean values which are computed by iterative MAP adaptations, see section 2.4.4. On each iteration zero and first order Baum-Welch statistics $\vec{N}_c, \mathbf{F}_c$ are influencing the deterministic computation of a samples supervector $\vec{\mu}_\omega$. Hence, the initial Baum-Welch statistics are already representative for a supervector. Further, the centered first order Baum-Welch statistic $\mathbf{F}_{c,centered}$ of the initial MAP iteration represents the supervector – UBM offset, see section 2.4.2.1. Thus, a conceptional iterative $\mathbf{T}$ matrix training paradigm can be defined by the least square error between the UBM offset and its decomposition terms from eq. 73:

$$\arg_{\mathbf{T},\vec{i}(\chi)} \min ||\vec{F}(\omega) - \mathbf{T}\,\vec{i}(\chi)||^2 \tag{74}$$

where $\vec{F}(\omega)$ denotes the supervector-alike concatenation of centered first order Baum-Welch statistic $\mathbf{F}_{c,centered}$ of all GMM components c. Where eq. 74 can be interpreted as a minimum divergence and a maximum likelihood, respectively. An initial total variability matrix can be set randomly.

As the speech signal's MFCC features are assumed to be Gaussian distributed, a speaker's i-vectors are assumed to be Gaussian distributed as well. An universal i-vector variance $l_\mathbf{T}$ is influenced by the total variability matrix $\mathbf{T}$, the diagonal UBM covariance $\boldsymbol{\Sigma}_{\text{UBM}}$, and the zero order Baum-Welch statistics $\vec{N}_c$ which are concatenated and expanded to a diagonal matrix $\vec{N}(\omega)$. The zero-order Baum-Welch statistics represents the posterior probability of a UBM component or likewise the component's influence on a speaker GMM, hence the UBM covariance needs to be weighted by $\vec{N}(\omega)$. By using the total variability matrix and its transposed form, the variance effects can be mapped into the i-vector space:

$$l_\mathbf{T}(\omega) = \mathfrak{I} + \mathbf{T}'\boldsymbol{\Sigma}_{\text{UBM}}^{-1}\vec{N}(\omega)\mathbf{T} \tag{75}$$

where $\mathfrak{I}$ denotes the identity matrix. The expected value of a normal distributed i-vector is then computed by two standardised moments, first by normalising the centered first order Baum-Welch statistics $\vec{F}(\omega)$ with the UBM covariance and mapping it into the i-vector space, second by normalising the mapped value using the i-vector variance:

$$\mathsf{E}[\vec{i}(\chi)] = l_\mathbf{T}(\omega)^{-1} \circ \mathbf{T}'\boldsymbol{\Sigma}_{\text{UBM}}^{-1}\vec{F}(\omega). \tag{76}$$

where $\circ$ denotes the *Hadamard product*.

During training of the total variability matrix all samples $\omega \in \Omega$ are treated as being produced by different speakers, hence all within-

and between-variabilities are modelled as well. Total variability matrices are re-estimated by a maximum likelihood adaptation of the modelled variabilities with respect to the expected variabilities. Therefore, two accumulators are defined: $\mathfrak{A}_c$ representing the component-wise i-vector covariance weighted by zero order Baum-Welch statistics over all samples, and $\mathfrak{C}$ representing the variance between centered first order Baum-Welch statistics and i-vector means:

$$\mathfrak{A}_c = \sum_{\omega \in \Omega} \vec{N}_c \, l_{\mathbf{T}}(\omega)^{-1}, \tag{77}$$

$$\mathfrak{C} = \begin{bmatrix} \mathfrak{C}_1 \\ \vdots \\ \mathfrak{C}_C \end{bmatrix} = \sum_{\omega \in \Omega} \vec{F}(\omega) \, E[\vec{i}(\chi)]'. \tag{78}$$

The total variability matrix **T** can be computed block-wise by solving

$$\mathbf{T}_c \, \mathfrak{A}_c = \mathfrak{C} \tag{79}$$

which is according to Kenny [81] a simplified eigen-analysis within the PCA case of re-estimating an eigenvoice model. Fig. 26 shows an exemplary total variability matrix of a short-duration, close-context development set containing German digits from zero to nine. By interpreting the total variability matrix as an offset mapping, some component offsets are more important to all i-vector elements than others (zero weight), and i-vector elements may depend more strong on certain GMM components rather than on all.

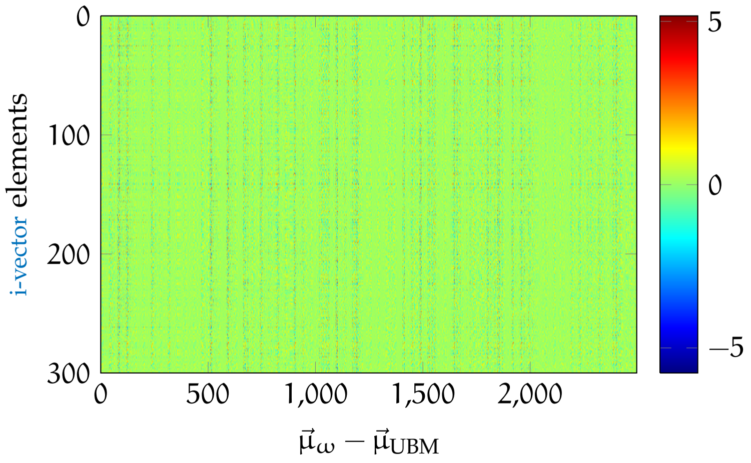


Figure 26: Total variability matrix example: mapping the supervector – UBM offset $\vec{\mu}_\omega - \vec{\mu}_{\mathrm{UBM}}$ (64 components of 39 MFCCs = 2 496 dimensions) to 300 dimensional i-vectors

3.4 IDENTITY VECTOR SYSTEMS

identity-vectors (i-vectors) can be directly extracted from a sample by processing the Baum-Welch statistics, so that the expected i-vector is estimated as in eq. 76 [7]:

$$\vec{\iota}(\chi) = (\mathcal{I} + \mathbf{T}'\boldsymbol{\Sigma}_{\mathrm{UBM}}^{-1}\vec{N}(\omega)\mathbf{T})^{-1} \circ \mathbf{T}'\boldsymbol{\Sigma}_{\mathrm{UBM}}^{-1}\vec{F}(\omega) \tag{80}$$

where the extracted i-vector $\vec{\iota}(\omega)$ represents speaker-characteristic supervector – UBM offsets. An i-vector-based speaker space can be created by using UBM development samples:

- centering i-vectors by a-priori observed means $\vec{\mu}_{DevSet}$ leads to an averaged supervector – UBM offset representation where opposite directional vectors implicate polarised speaker sub-spaces which could have more similar directions before centering,
- then i-vectors are *transformed into a spherically symmetric density by a linear whitening transformation learned from data samples* [82], see section 2.4.7, so that i-vector elements are decorrelated by a whitening matrix $\mathbf{W}$ and hence span a vector space,
- further, data shifts are compensated by length-normalisation where data set mismatches between development and evaluation data may cause huge differences on lengths of similar directional i-vectors of the same speaker [82].

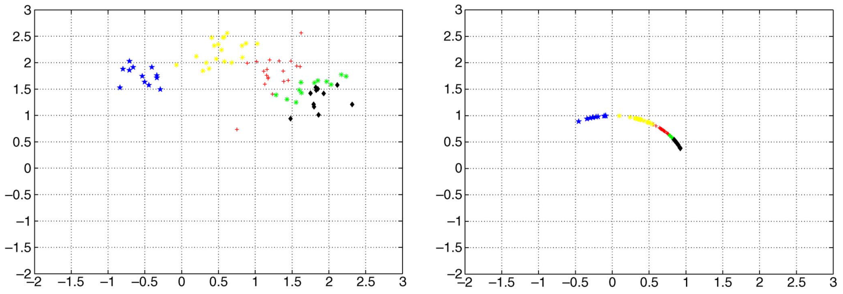
Raw i-vectors $\vec{\iota}_{\mathrm{raw}}$ are transformed into unit spherical i-vectors $\vec{\iota}_{\mathrm{unit}}$ by applying the following equation:

$$\vec{\iota}_{\mathrm{unit}} = \frac{(\vec{\iota}_{\mathrm{raw}} - \vec{\mu}_{DevSet})\mathbf{W}}{||(\vec{\iota}_{\mathrm{raw}} - \vec{\mu}_{DevSet})\mathbf{W}||} \tag{81}$$

where further in this thesis $\vec{\iota} = \vec{\iota}_{\mathrm{unit}}$ will be denoted to ease notations. These i-vectors are state-of-the-art features [45, 46] which can be used as templates $\vec{\iota}(\chi)$ of speakers χ and probes $\vec{\iota}(\omega)$ of verification samples ω as well. Fig. 27 illustrates exemplary i-vector spaces before and after the unit-sphere transformation where the spherical speaker placement promises a very good speaker-separability in terms of biometric recognition.

3.4.1 *Biometric enrolment and verification*

Biometric speaker recognition systems comprise enrolment, re-enrolment, and verification processes. Speaker reference templates are created as mentioned by extracting an i-vector from an enrolment sample. For re-enrolments Ferrer et al. [83] proposed the average template of all enrolment i-vectors, since the average i-vector is assumed to be more robust towards within-speaker variabilities, compare fig. 27.



(a) Raw i-vectors

(b) Unit i-vectors

Figure 27: i-vector space before and after unit-sphere transformation according to [7] where both axis denote two different i-vector elements for five different coloured speakers

During verification processes i-vector features are extracted and used as probes. i-vectors can be compared by e. g. the cosine similarity between them, which can be computed as the length normalised dot product of template $\vec{i}_t$ and probe $\vec{i}_p$ i-vectors:

$$S_{\cos} = \frac{\vec{i}_t \cdot \vec{i}_p}{||\vec{i}_t||\,||\vec{i}_p||}. \tag{82}$$

In this thesis focus is placed on the simple cosine distance as i-vector comparator. However, during the last years *Gaussian Probabilistic Linear Discriminant Analysis* (G-PLDA) became popular among speaker recognition for scoring the likelihood of two i-vectors[14].

3.4.2 *Score normalisation*

Score normalisations are applied to augment comparison scores with additional information e. g., about an enrolled reference or a current verification probe. For the purpose of applying standard score normalisation methods by preserving the symmetry between i-vectors, Kenny [24] introduced the *spherical normalisation* (s-norm). S-norm relies on the two standard normalisations of zero and test score normalisation which are both computed similar by normalising a score S to S' by mean and variance μ, σ of observed score distributions:

$$S' = \frac{S - \mu}{\sigma}. \tag{83}$$

The *zero score normalisation* (z-norm) computes the score mean μ_3 and standard deviation σ_3 of a template i-vector compared against an

[14] PLDA is a a special case of both, LDA and JFA: *the relationship between PLDA and standard LDA is analagous to that between factor analysis and principal component analysis* [84]. G-PLDA is a Gaussian PLDA variant: i-vectors are assumed to be Gaussian distributed, such that each i-vector emission has a posteriori probability $\mathcal{N}(\vec{\mu}_{\text{G-PLDA}} + \mathbf{W}\vec{h} + \mathbf{B}\vec{g}, \mathbf{\Sigma}_{\text{G-PLDA}})$ which is influenced by prior-trained speaker-within and -between variances $\mathbf{W}, \mathbf{B}$ resulting in JFA-like decomposition of characteristic speaker- and noise-factors $\vec{h}, \vec{g}$ [24, 84, 85, 86, 87].

i-vector collection $\mathfrak{Z}$, and the *test score normalisation* (t-norm) compares similar parameters $\mu_{\mathfrak{T}}, \sigma_{\mathfrak{T}}$ of a probe i-vector against an i-vector collection $\mathfrak{T}$. Hence, a verification score S can be normalised by centering impostor scores having unit variance by known impostor score distributions with respect to a template i-vector and of a probe i-vector as if it was an impostor i-vector,

$$S' = \frac{1}{2}\left(\frac{S - \mu_{\mathfrak{Z}}}{\sigma_{\mathfrak{Z}}} + \frac{S - \mu_{\mathfrak{T}}}{\sigma_{\mathfrak{T}}}\right). \tag{84}$$

However, the s-norm can be computed more robust by adaptively selecting collection subsets of $\mathfrak{Z}, \mathfrak{T}$ which is referred to as adaptive spherical score normalisation (AS-norm) [50, 74, 75]. The AS-normalised score S' differs from s-norm by the scores which are used to compute the z/t-statistics: rather than using all scores, only the most competitive scores (*e.g.* top-100) are applied to model according speaker cohorts. Dehak *et al.* [64] applied the AS-norm on i-vectors and showed that the score normalisation can already be applied on comparison-level as a normalised cosine scoring for a template-probe i-vector $\vec{\imath}_t, \vec{\imath}_p$ comparison:

$$S(\vec{\imath}_t, \vec{\imath}_p) = \frac{(\vec{\imath}_t - \vec{\imath}_{\mu_{\mathfrak{Z}}})^T (\vec{\imath}_p - \vec{\imath}_{\mu_{\mathfrak{T}}})}{\|\boldsymbol{\Sigma}_{\mathfrak{Z}} \vec{\imath}_t\| \, \|\boldsymbol{\Sigma}_{\mathfrak{T}} \vec{\imath}_p\|} \tag{85}$$

where $\vec{\imath}_{\mu_{\mathfrak{Z}}}, \vec{\imath}_{\mu_{\mathfrak{T}}}$ denote mean i-vectors of z- and t-norm collection sets and $\boldsymbol{\Sigma}_{\mathfrak{Z}}, \boldsymbol{\Sigma}_{\mathfrak{T}}$ are according diagonal covariance matrices. However, the emphasis of this thesis is placed on AS-norm, since the normalised cosine is an optimisation step.

3.4.3 *System fusion*

Speaker verification systems can be fused on different levels depending on their comparators. HMM, GMM, and i-vector-based verification systems can be fused on the score-level domain by e. g., logistic regression. Logistic regression fuses the scores of subsystem $\mathfrak{s} = 1, \ldots, \mathfrak{S}$ into one score S' by a weighted linear combination of all subsystem scores $S_{\mathfrak{s}}$ [40]. A score fusion formula might take a general score offset a into account as well as additional quality-based information which can be briefly denoted by a quality function $\mathfrak{Q}$ [25, 40, 88]:

$$S'(\vec{\imath}_t, \vec{\imath}_p) = a + \sum_{\mathfrak{s} \in \mathfrak{S}} b_{\mathfrak{s}} S_{\mathfrak{s}}(\vec{\imath}_t, \vec{\imath}_p) + c\mathfrak{Q}(\vec{\imath}_t, \vec{\imath}_p) \tag{86}$$

where $b_{\mathfrak{s}}, c$ are weights. The parameters $a, b_{\mathfrak{s}}, c$ need to be determined by logistic regression. Logistic regression maximises the likelihood of a prediction model to perform well [20]. Thereby, the cross-entropy error $\mathcal{H}$ is minimised which in the case of speaker recognition can be the C_{llr} metric [40, 41], see section 2.3.5.

Logistic regression is an iterative learning method which uses the gradient descent: for each weight configuration $\vec{w} = \langle a, b_{\mathfrak{s}}, c\rangle$ a cross-entropy-based gradient $\nabla\mathcal{H}$ is calculated on N training scores with labels y_n by [20]:

$$\nabla\mathcal{H}(\vec{w}) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n S_n}{1+e^{y_n \vec{w}^T S_n}}. \tag{87}$$

The weight configuration is updated by $\vec{w} = \vec{w} - \nu\nabla\mathcal{H}(\vec{w})$ where ν is a learning rate coefficient which is usually set to a small positive real[15] [20]. A logistic regression model is found if the gradient descent is zero or a maximum number of iterations e. g., 100, is reached, see algorithm 1.

```
Data: Initial weights $\vec{w}_0 = \langle a, b_{\mathfrak{s}}, c\rangle = \vec{0}$
Result: Final weights $\vec{w}$
while i=1,...,100 do
    Compute $\nabla\mathcal{H}(\vec{w}_{i-1})$;
    if $\nabla\mathcal{H}(\vec{w}_{i-1}) == \vec{0}$ then
        break;
    end
    Update weights $\vec{w}_i = \vec{w}_{i-1} - \nu\nabla\mathcal{H}(\vec{w}_{i-1})$;
end
```

Algorithm 1: Logistic regression, see Abu-Mostafa et al. [20]

Further, Ferrer et al. [83, 89, 90] proposed a feature-level i-vector fusion where i-vectors extracted on different speech signal features, i. e. prosodic polynomial contours[16] (ProsPols) [83], are merged to a *grand i-vector* on which LDA can be performed in order to obtain low-dimensional i-vectors and removing information redundancy. Score-level and grand i-vector speaker verification system fusions are compared in fig. 28 with respect to fig. 3. In this thesis focus is placed on score-level subsystem fusion, because different comparator based subsystems are addressed.

3.5 SUMMARY

In this chapter speaker recognition methodologies were presented, that motivated by human anatomy and articulatory motion. Speech is transmitted by changes in air pressure which is influenced by e. g., a humans fundamental frequency characterising their voice by the glottal pulse rate that depends on the shape of a human's vocal fold. Besides biological characteristics, the human voice also has behavioural

[15] E. g., 0.1, 0.01, or 0.001.

[16] ProsPols are *Legendre polynomial approximations of order 5 of pitch and energy signals* over a region of 20 segments [83].

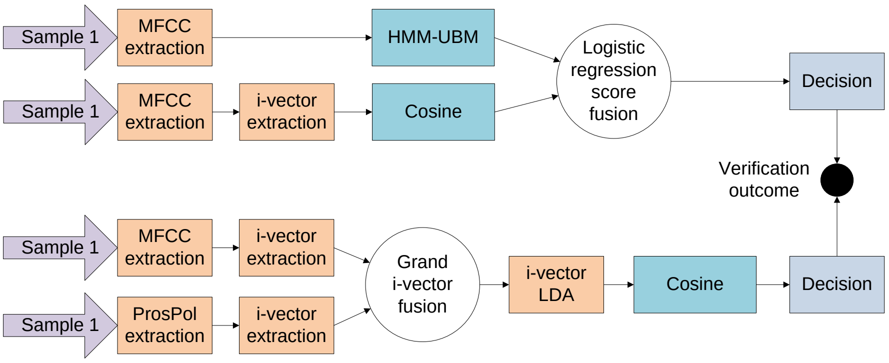


Figure 28: i-vector system fusions on score- (top) and feature-level (bottom)

characteristics which depend on the person's cultural background as well as the spoken language or dialect as well. In order to extract characteristic features, VAD is applied, such that non-speech signals are discarded before any feature is extracted. Fig. 29 summarises speech signal processing from VAD processing of the sample, MFCC feature extraction, to UBM-based speaker template, reference, and probe creations with respect to the design of a general biometric system, see fig. 2 in section 2.2.

A basic MFCC feature extraction transforms the time-domain raw-values into the frequency domain by Fourier transform, so that coefficients describing speaker-dependant frequency bandwidths can be computed. MFCCs are obtained by transforming the coefficients back into time-domain using the inverse discrete cosine transformation. However, speech variations also depend on the spoken content which has basic units, the phones or phonemes, respectively. By extracting features from short-time segments, detailed information about the articulatory motions of speakers can be computed.

Further, signal processing also computes velocity and acceleration coefficients in order to augment the short-term MFCCs. The extracted features are then normalised by feature warping, so that within-sample variations are minimised: the observed MFCCs are mapped onto a standard distribution according to their value within an e. g., 300 features sequence. MFCCs are referred to as being independent and Gaussian distributed.

An universal feature space of MFCCs is modelled by GMMs where a speaker-independent cluster of the acoustical space is referred to as UBM. By modelling the whole acoustical space using GMMs, speaker recognition becomes text-independent, since features can be scored without knowing the spoken context surrounding them. In contrast, HMM-based UBMs are text-dependant, since they model e. g., a sequence of words or phones. However, by knowing the spoken text, HMMs modelling text-parts can be combined and thus, HMMs can be used for text-independent speaker verification under higher computational effort as well. All further speaker recognition processing steps

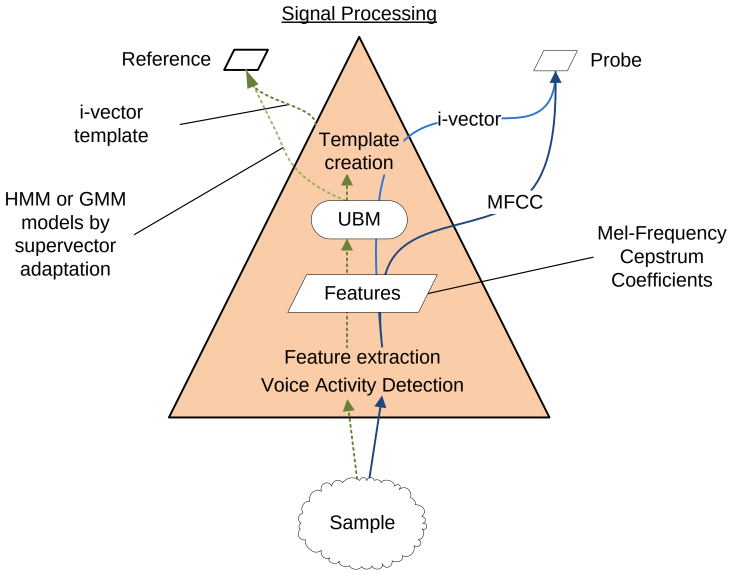


Figure 29: Overview on biometric speech signal processing

depend on the quality of this UBM. In order to reduce over-fitting of too frequently occurring features, IFS was motivated which selects only those features which are contributing new information towards an UBM.

Speakers are assumed to create subspaces within the acoustical space, hence biometric speaker verification scores the probability of a probe belonging more to the claimed speaker's subspace, rather than to the UBM space. Thereby, subspaces of speakers are estimated by MAP adapting the UBM means using their enrolment samples. Hence, the UBM means form speaker templates which are referred to as supervectors. In GMM-UBM or HMM-UBM systems probe samples are scores against the UBM and the supervector-based speaker model. Where supervectors are also strongly influenced by environmental noise and the articulated phones.

identity-vectors (i-vectors) were motivated in order to extract only speaker-characteristic features or likewise principal speaker components which do not depend on other speech signal influences, such as channel noise or the phonetic content. Therefore, supervector – UBM offsets were analysed with respect to all variabilities within a development set. Total variability matrices are trained in order to decompose these offsets into i-vectors by a minimum divergence between the supervector – UBM offset and it's total variability/i-vector representation among a development set. Hence, the total variability matrix can be seen as a mapping from supervector offsets to i-vectors. Fig. 30 gives an overview on HMM-UBM, GMM-UBM, and cosine i-vector comparators.

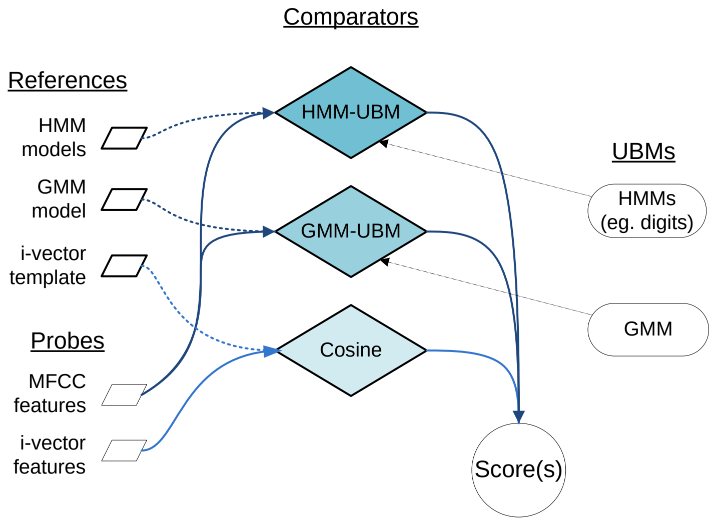


Figure 30: Overview on speaker verification comparators

Further i-vector processing spans a spherical unit space of i-vectors by space-centering, axis-decorrelation due to whitening, and vector length normalisation. An i-vector scoring can then be performed by the simple cosine distance as similarity metric. In order to augment verification scores with knowledge among the development set, score normalisations can be applied, where the adaptive spherical normalisation (AS-norm) applies zero- and test-normalisations (z-, t-norm). While z-norm estimates the distribution of impostor scores on the reference of a subject, t-norm estimates the distribution of impostor scores using a current probe sample. By taking only the top-k scores for each distribution estimation into account, the normalisations are supposed to become adaptively and thus more robust. The spherical shape of i-vectors can be preserved by simply averaging z- and t-normalised scores. Speaker verification systems can be fused e. g., on score-level by a-priori training a linear regression model such that the verification cross-entropy C_{llr} is minimised.

Part II

SHORT AND VARIANT DURATION SCENARIOS: METHODOLOGIES, FRAMEWORK, AND EVALUATION

4 SHORT DURATION AND VARIANT DURATION SPEAKER VERIFICATION

The performance of speaker verification systems strongly depends on the speech sample durations [25, 91, 92]. Long duration samples deliver much more information about a speaker in terms of the extraction of sufficient statistics compared to short duration samples. Short duration samples comprise only few features that are known to result in a poor speaker representation [17]. Though, the rejection of impostors might be possible after few seconds of speech [93, 94], investigations are needed for accurately accepting genuines as well.

This thesis places emphasis on short duration speaker verification in close context scenarios, and on variant duration speaker verification which is a current NIST research task on speaker recognition [46, 69, 95]. This chapter provides an overview on related literature on short and variant duration research for which deductive hypothesis will be pointed out that will be examined within the thesis evaluations.

4.1 SHORT DURATION SPEAKER VERIFICATION

Short duration samples are relevant for e. g. pass-phrase-based industry scenarios [96], and on phonetically-closed forensic scenarios [17, 97]. Fatima and Zheng [97] proposed a research agenda for short duration speaker recognition where they differentiated between speaker modelling approaches e. g., HMM-UBM, GMM-UBM, JFA and i-vector systems. For the purpose of obtaining more information about short duration speech, they supposed to analyse phonetic categories using the known approaches, so that e. g., vowel, nasal, and consonant phones, are compared. This approach seems reasonable on very short duration samples, such as short yells for help which are given e. g., in the shooting case of *Trayvon Martin* [2]. But the speech data can just be analysed without taking between-phone variations into account, since there might be no comparable data.

Zhang et al. [98] examined phone-based GMM-UBM systems which only use the most present phones within samples, so that sufficient statistics can be computed for the top-k phones of each sample and used for verifications. Thereby, phone GMMs are referred to be more accurate on according phone-based speech data than text-independent GMMs are. They emphasised samples having less than two seconds of speech.

However, on automated scenarios comprising e. g., more than 15 phones within five seconds (five digited pass-phrases), additional computational effort will arise in terms of *accurately* extracting phone dependent speech subsequences. For the purpose of examining industry scenarios relying on pass-phrases and in order to avoid additional mismatch risks of speech subsequence alignments by e. g. HMMs. Phone-based model techniques will not be examined in this thesis as much as the i-vector approach. Thus, emphasise is placed on text- and language-independent approaches. Larcher et al. [96] evaluated i-vector systems in scenarios with phonetically constrained contexts in terms of commands and passphrases. By reducing the modelled phonetic content, their systems became more accurate on according evaluation data.

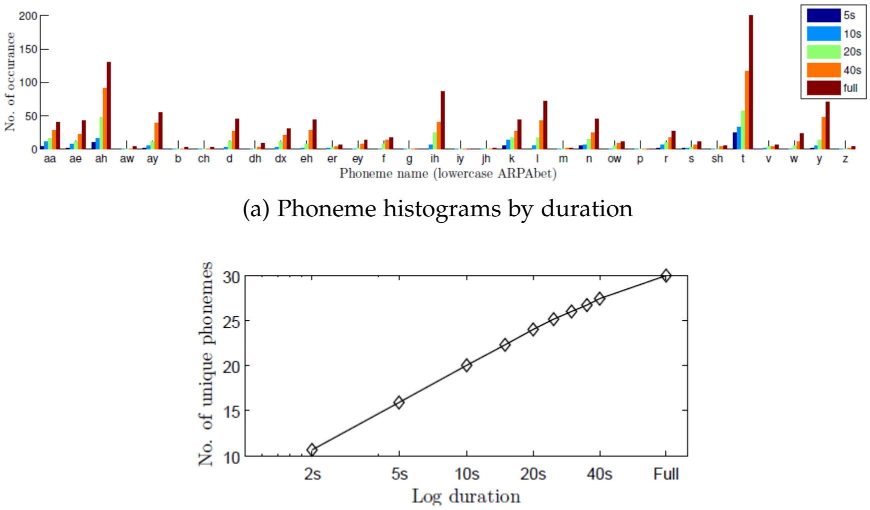
4.2 HYPOTHESIS ON SHORT DURATION SCENARIOS

Fatima and Zheng [97], Zhang et al. [98], and Larcher et al. [96] are following the same approach: if the phonetic content is a close set, then verification accuracy can be improved by modelling only the expected content, thus under-fitting of the UBMs and speaker models can be avoided. Hence, if short duration scenarios are aimed, then *divide and conquer* algorithms with respect to the phonetic content and channel effects are applicable: computational efforts will not exceed real-time requirements, though enrolment samples need to have verification sample alike shapes, which could be achieved by e. g., multi-session enrolments containing different subsequences of the expected spoken phrases within verification trials. However, due to the a-priori knowledge of possible verification samples, the usage for (full) text- and language-independent applications might not be given, since either the phonetic content is a closed-set, hence the system configuration is fix, or according intensive enrolment sessions will be necessary in order to avoid poor comparison quality due to context mismatches.

Concluding hypothesis: *for the purpose of robustly comparing short speech patterns, short duration scenarios need to be separated from each other and to be solved on similar phonetic contexts and on similar signal quality shapes such as commands, passphrases, digits or phones, and phonemes which comprise industry and forensic terms, respectively.*

4.3 METHODOLOGIES ON VARIANT SAMPLE DURATIONS

Varying sample durations cause huge gaps between speech samples: depending on the phonetic content supervectors vary in terms of over- and under-presented statistics of the UBM components. This effect was examined by Hasan et al. [88] and is referred to as *acoustic holes*. Acoustic holes can be explained as the miss or under-representation of phonemes in short duration samples which are much more represented in mid-term (20–40s) and long term samples (> 40s, full), see fig. 31. Further, Hasan et al. [88] showed that the dependency of unique phone occurrences to the sample duration is logarithmic, see fig. 31b.



(a) Phoneme histograms by duration

(b) Duration variance causing acoustic holes

Figure 31: Illustration of acoustic holes according to [88]

By training i-vector systems using variant-duration samples e. g., due to random sample truncations, performance gains were reported [88, 92]. However, Vogt et al. [99] showed that the trace[1] of eigenvoice matrices trained by different sample lengths increase by decreasing the sample durations: from 105.7 on > 40s samples to 329.8 on < 10s samples of the 2005 NIST SRE data. Which illustrates that first there are totally different dependencies within the factor decomposition matrices and second that on shorter samples the factor influence of the components grows a lot, up to 312%. Hence, factor analyses would need to be performed depending on the constellation of enrolment and verification sample durations which is computational not reasonable, thus a duration-invariant i-vector extractor will need post-processing compensations of duration mismatches.

Mandasari et al. [25] and Hasan et al. [88] further examined the use of Quality Model Functions (QMFs) which are applied as in the logistic regression case of system fusions, see section 3.4.3. Thereby, the QMFs rely on enrolment and verification sample durations d_e, d_v, such that the logarithmic observations from fig. 31 are preserved. A possible QMF can be denoted by the duration LLR's absolute value [25, 88]:

$$Q(d_e, d_v) = \left| \log \frac{d_e}{d_v} \right| \tag{88}$$

[1] The trace of a matrix is the sum of it's diagonals.

which is a log-symmetric distance metric. According to Mandasari et al. [25] this score calibration method is able to yield gains compared to a system calibration applying only logistic regression as in eq. 86.

Further, Mandasari et al. [17, 25] evaluated the performance of i-vector systems with respect to enrolment-verification duration groups where better performances were observed on > 40s duration enrolment samples. Improvements were reported on by normalising the i-vectors as described in eq. 85 using the complete development set [17], and by applying logistic regression with respect to duration groups of < 5s, < 10s, < 20s, 40s, and > 40s (full) samples [25] which were chosen according to the logarithmic dependency observation regarding to the acoustic holes [88]. In comparison to the QMF calibrations, the duration-matched score calibrations were reported to averagely yield less entropy in terms of C_{llr}. QMFs comprise three to five free calibration parameters and the duration-matched calibration comprises 50 free parameters [25] that need to be determined by logistic regression.

However, by setting a shared scaling parameter – in terms of eq. 86 the weight b is shared among all 25 duration group constellations – Mandasari et al. [25] observed a saddle-plane shaped distribution of the free offset parameters, see fig. 32. Where the saddle-plane shape seems to be symmetric by log-durations, and edge mismatches might have been occurred due to hard-decisioned duration group memberships e. g., 0–5s belonging to a 5s duration group. Though, soft-memberships as mentioned by eq. 88 seem not to have this hard-decision boundaries. Further, on full enrolment i-vectors, which will be emphasised in this thesis, the offset parameters are distributed in a similar logarithmic manner as the acoustic holes to the speech signal duration reported in fig. 31b.

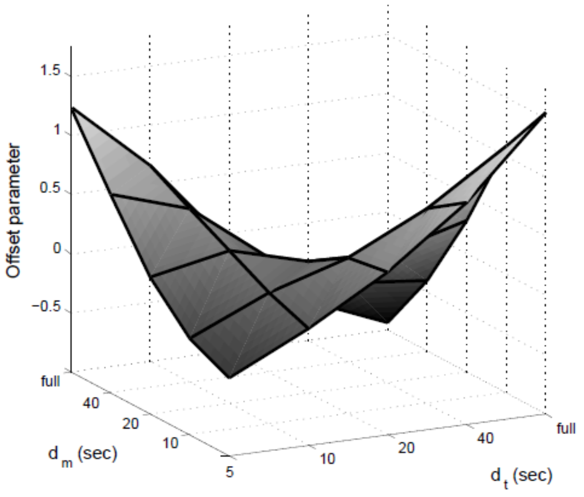


Figure 32: Calibration parameter distributed in saddle-plane shape on log-durational axes according to [25] where $d_m = d_e, d_t = d_v$

4.4 hypothesis on duration variance and subspaces

Acoustic holes are the issue to deal with focusing on duration variance. They are shown to logarithmically depend on the sample duration [88]. Techniques compensating performance break-downs rely on score calibration models such as QMFs [25, 88], on training the recognition system with data of various sample durations [88, 92], or on different i-vector extraction/post-processing [17, 99]. Systems whose performances strongly relies on their training data seem not high-convenient, and total variability matrices trained according to various duration constellations seem also not to be sufficient for facing unknown speech data.

However, Vogt et al. [99] showed the existence of duration-depending subspaces within an eigenvoice model, hence this effect needs to be compensated as well in order to compare same spaced unit i-vectors. Further, Mandasari et al. [25] presented the advantage of duration-matched compensation approaches in comparison to generalised QMF functions which might yield more gains on same-spaced i-vectors. Though, for the purpose of keeping the compensation of acoustic holes as simple as possible and providing sample-adaptive acoustic hole compensation methods, the training of a fix logistic regression parameter set seems to emphasise the compensation of other effects, too.

Concluding hypothesis: *the divide and conquer strategy of treating probe i-vectors by a-priori knowledge of duration-matching i-vector subspaces promises to dynamically compensate the effects of acoustic holes. This implies the i-vector relocation into a common subspace that can be realised by relocating template and probe i-vectors. However, the issue of acoustic holes or likewise quality mismatches is less about compensating the duration mismatches, it is more about covering any articulatory motion. Thereby, algorithms need to acknowledge that the produced speech within a subject's enrolment sample comprise very sparse enrolment data in comparison to the diversity of sounds which can be produced by humans. Further, the comparison algorithms need also to acknowledging the signal's unsteadiness in terms of the human voice as well as in terms of channel noise effects. That also is the compensation of duration mismatches which is a more feasible issue to emphasise research on.*

4.5 duration invariant adaptive score normalisation

In order to adaptively relocate template and probe i-vectors into a common subspace, techniques need to be established that can estimate the subspaces of each template and probe i-vectors. For this purpose compensations e. g., based on feature-level and score-level, are possible where this thesis emphasises on score-level compensations.

The AS-norm uses information about the score distribution of both, template and probe i-vectors, among a development set to normalise verification scores, see eq. 84. Thereby, the most similar development i-vectors are adaptively selected for estimating template/probe-characteristic i-vector subspace properties on the score-level domain. By selecting development set i-vectors according to the quality mismatches that are measured in terms of sample durations or likewise the impact of acoustic holes, a duration invariant adaptive score normalisation can be motivated as an extension of the standard AS-norm.

As previously mentioned, the presence of acoustic holes increases the entropy of shorter voice samples due to performance losses which motivates the construction of different i-vector sufficiency-classes.

In terms of duration as a quality metric, Q quality classes can be denoted as: $\mathfrak{Q} = \{\Lambda_0, \dots, \Lambda_Q\}$ representing i-vector sufficiency classes. Samples are then associated by their duration d_s to a sufficiency class Λ_c by the lowest log-duration distance,

$$\arg_{\Lambda_c} \min |\log(d_s) - \log(d_{\Lambda_c})|. \qquad (89)$$

4.5.1 *i-vector sufficiency classes*

In the proposed system duration-based groups are defined for the sufficiency classes where the number of quality classes is limited to $Q = 5$, i. e. obtained results can be directly compared to those reported in [25, 88]. By preserving comparability to the related work on short-duration speaker recognition, sufficiency classes are denoted according to the researches on acoustic holes of Hasan et al. [88] and Mandasari et al. [25]. The proposed sufficiency-classes are summarised in tab. 3 where Λ_{full} is intended to comprise all expected high-sufficient i-vectors which might cause non-optimal results, but preserves low-computation efforts.

Table 3: Sufficiency classes and corresponding durations

Sufficiency class	Duration
Λ_5	0–5 sec
Λ_{10}	5–10 sec
Λ_{20}	10–20 sec
Λ_{40}	20–40 sec
Λ_{full}	$\geqslant 40$ sec

Fig. 33 illustrates sufficiency-class dependencies with respect to a sample's duration according to eq. 89. The maximal neighbour distance of sufficiency-class within the denoted classes $\Lambda_{5\text{–}40}$ is $\log\sqrt{2}$ and $\log 2$ between two neighbour classes. Whereby, the classes Λ_5 and Λ_{full} comprise several quality classes e. g., of 1, 2, 80 and 160 second durations. In tab. 3 quality classes are labelled by the upper duration bound and samples are assigned to classes by the next highest duration class having a maximum distance of $\log 2$, besides $\Lambda_{5,\text{full}}$ which comprise extreme boundary quality classes.

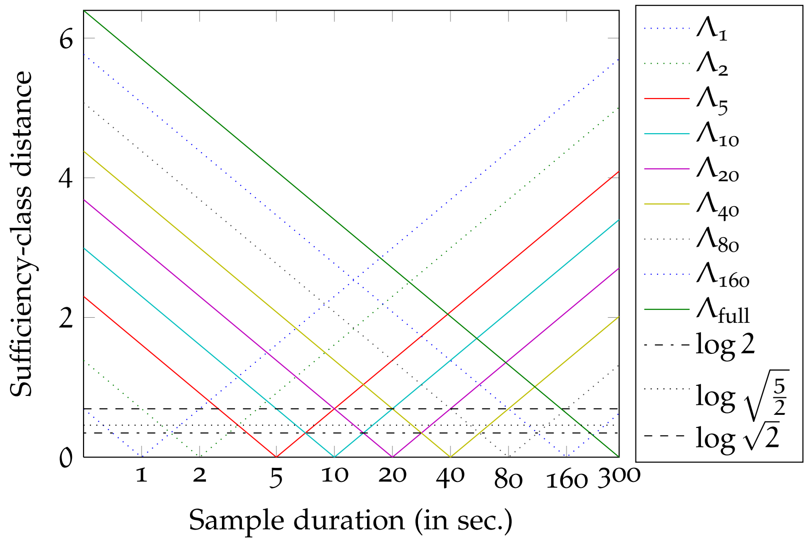


Figure 33: Duration-based filter for i-vector sufficiency-classes denoting different quality shapes of speaker-representation where summarised classes are dotted

4.5.2 *Parameter estimation*

For the z- and t-norm parameter AS-cohorts are pre-selected in the following manners:

- z-norm simulates impostor verifications on averaged enrolment templates $\vec{\imath}_t$, thus only $\mathfrak{Z}$ i-vectors will be used which belong to the same sufficiency class as the probe i-vector Λ_{d_p}:

$$\mathfrak{Z} = \{\vec{\imath}_{\Lambda_{d_p}} | \max_{\text{top100}} S(\vec{\imath}_t, \vec{\imath}_{\Lambda_{d_p}})\}, \tag{90}$$

- t-norm simulates impostor verifications comparing the probe i-vector $\vec{\imath}_p$ to development set templates, where enrolled speakers have full i-vectors, so that the vast majority of durations is greater than 60 seconds $\Lambda_{>60}$ e. g., only $\mathfrak{T}$ i-vectors will be used extracted from samples with longest durations:

$$\mathfrak{T} = \{\vec{\imath}_{\Lambda_{>60}} | \max_{\text{top100}} S(\vec{\imath}_p, \vec{\imath}_{\Lambda_{>60}})\}. \tag{91}$$

4.5.3 *Score estimation*

The proposed duration-adaptive extension of AS-norm normalises the scores according to eq. 84. By placing emphasis on duration-based sufficiency classes, recognitions are treated duration invariantly e. g., normalised scores are expected to be distributed without creating entropy due to duration mismatches. Further, an overall improvement can be expected, since scores of all sufficiency classes are normalised to more similar distributions of genuine and impostor scores. As a consequence, no additional entropy is expected, which could arise due to score-distribution mismatches, i. e. caused by fix across-classes thresholds. Fig. 34 illustrates how variant-duration samples will be processed by the proposed duration invariant AS-norm extension. Further, please find our submitted *Speaker Odyssey 2014* paper towards duration invariant AS-norm in appendix D.

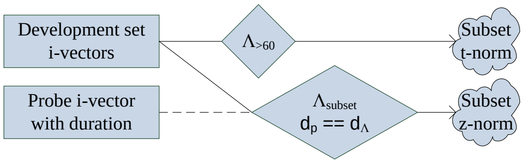


Figure 34: Duration-based normalisation subsets

5 SPEAKER VERIFICATION FRAMEWORK

For the purpose of designing a biometric speaker verification framework, this chapter entitles research requirements, concludes a system design that is based on ISO/IEC 19575-1 [26], and states a prototype framework's implementation. In the following sections requirements are named and then marked by (a-k). For the purpose of demonstrating where and how the requirements are covered by the framework design, the design description will refer to these marks. The last section of this chapter will explain a basic implementation of the proposed framework design.

5.1 REQUIREMENTS

In research terms, a speaker verification framework should comprise state-of-the-art recognition techniques such as the GMMs and i-vector approaches (a). Further, in more general terms, a framework should be easy to extend with other recognition methods (b) or further pre- and post-processing methods (c) and should include third-party recognition results as well (d) for system fusions. Since research may place focus on recognition parameters like the UBM size in terms of components, an according framework should give the opportunity to easily store and load data with respect to a system's configuration. Within configuration changes the according data should be replaced if existing or otherwise collected (e). For the purpose of providing well-defined framework processing, each parameter configuration requires an according implementation (f).

A speaker verification framework is intended to process existing speech samples[1] (g) and verify enrolled subjects by biometric pattern recognition, see chapters 2, 3. Thereby, industry-applicable subject use cases as enrolment, re-enrolment, and verification need to be concerned (h) as well as performance evaluation use cases where enrolment and verification processes are performed on a recognition system, such that performance metrics, as mentioned in section 2.5, can be calculated (i). Both use cases differ in terms of processing time: the subject-based use cases require single-processing while the evaluation case aims at fast computations (j) e. g., through parallel processing or multi-threading.

Further, the application-scope or evaluation scenario of GMM-based speaker recognition relies on a system's development database e. g., UBMs modelled by digit development speech corpora are more ap-

[1] Speech samples are expected to be in raw pulse-code modulation format.

plicable to digit scenarios than to open-context scenarios. Hence, a general framework needs to differ between several development sets, and various application or evaluation scenarios (k).

5.2 FRAMEWORK DESIGN

Following up the ISO/IEC 19575-1 general biometric system design [26], compare fig. 2, a speaker recognition framework needs to separate signal, comparison, and decision making processing.

5.2.1 *Biometric system processing*

Speaker recognition signal processing comprises reading raw speech samples (g), VAD, MFCC feature extraction, and feature warping, compare chapter 3. Where comparators rely on different features e. g., GMM-UBM systems process MFCCs as features, while i-vector systems are based on UBM Baum-Welch statistics (a), see sections 3.3, 3.4. Hence, signal processing might already use GMMs as shown in fig. 29 (a). Also, systems may differ in decision making e. g., in terms of applied score normalisation techniques, system fusions, or score calibrations (a), see sections 2.3, 3.4, 4.5. Therefore, Object-oriented Programming (OOP) provides class and type enumeration-based design patterns which benefit from the modularity of OOP: by providing a system class with abstract functions many requirements can be fullfilled where abstract functions are:

- `float** features := extract_features(sample) (c),`
- `void enrol(identity, features) (a, b),`
- `float* probe := create_probe(sample) (a, b)`
- `float score := compare(reference, probe) (a, b),`
- `float score := normalise(reference, probe, score) (c).`

Where `float*` shall indicate a vector of floats, and the abstract functions may use system depending parameters which are denoted as attributes. Such attributes can comprise e. g., feature types (VADs, MFCCs, feature warping), comparator parameters (UBM components, i-vector factors), and score post-processing methods (AS-norm, fusion, calibration) (a-d). All attribute variations can be defined by enumerations and in order to preserve well-defined processing, implementations need to be done according to each enumeration value (f).

Biometric systems are implemented by inheriting the abstract system class and defining the abstract methods by existing modules or using external third-party software. Thereby, a mocking inherited dummy system class provides the opportunity of only evaluating the results of a third-party speaker recognition system `(d)` whereby result file exchange formats need to be defined as well[2]. By defining mocking inherited dummy classes for each processing variant, the framework will also preserve the opportunity to include third-party software on every processing stage `(d)`.

For the purpose of preserving configuration variability and *hot data swapping*, each class has save and read functions where a class' attribute values are stored into Hierarchical Data Format V5 (HDF5) file databases[3] `(e)`.

An overview of the adapted ISO/IEC 19575-1 general biometric system design for speaker verification is given in fig. 35. Data of each processing step is saved into according databases if not existing and can be loaded if existing e. g., VAD samples, MFCC features, UBMs, comparison scores, and normalised scores. The evaluation of a system is then performed on the according score database.

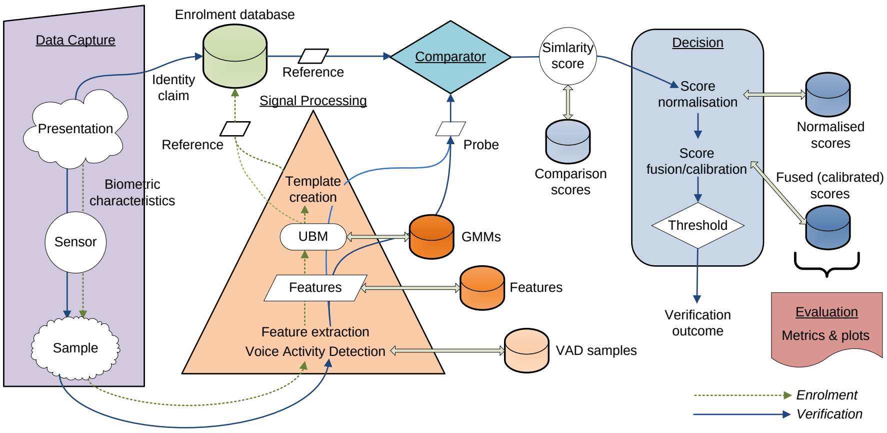


Figure 35: Design of the implemented speaker recognition framework

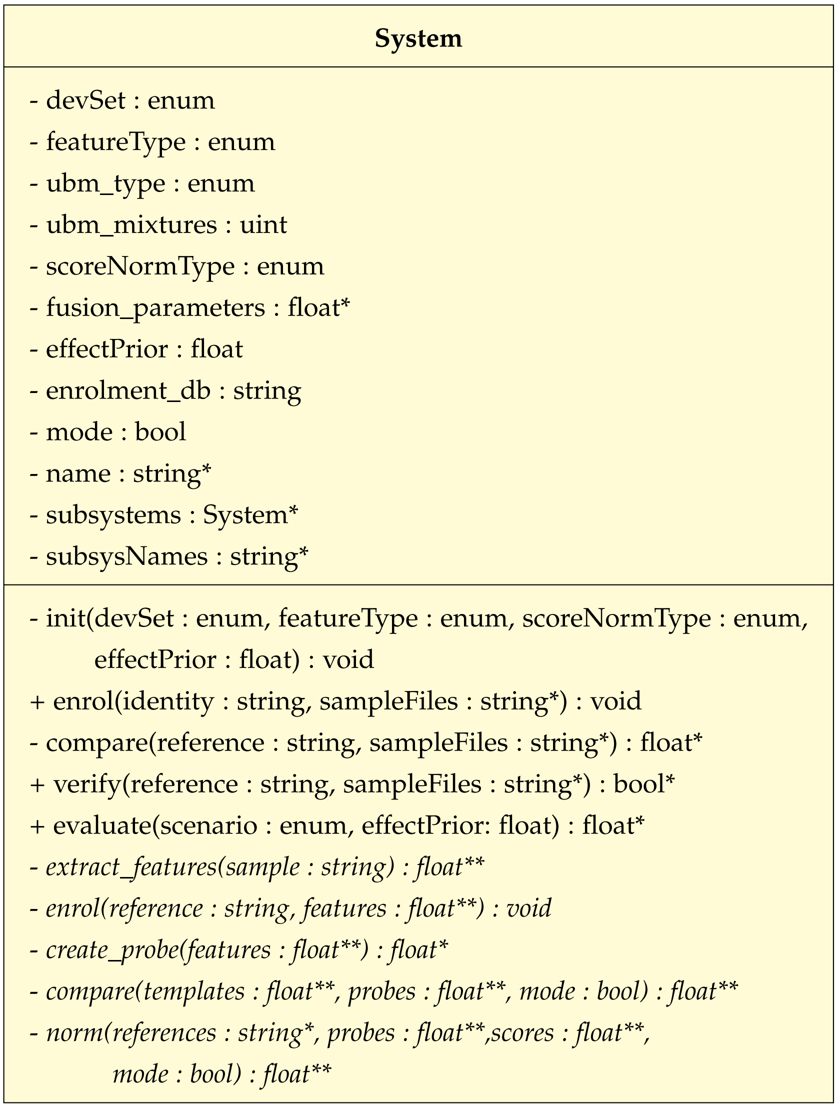
5.2.2 *Application and research modes*

For the purpose of using the framework in industry application and research scenarios, single calculated scores need to be returned as the framework's output such that a binary decision can be made by using thresholds, and evaluation scenario depending score mass computations need to be performed, so that the resulting performance of a system can be measured by them. Thereby, use of the application can be seen as a sub-part of the evaluation case where both modes can be separated by OOP cased interface methods `(h, i)`. Fig. 36 presents a basic class design for the abstract *System* class: all (im-

[2] E. g., by a comma-separated value (CSV) format of the shape: `speaker, claim, score, sample, transcriptions`.

[3] Hierarchical Data Format, a Matlab-common binary data storage type for large-scaled data in scientific purposes, see `http://www.hdfgroup.org/HDF5` (last viewed on 09.03.2014).

portant) parameters representing the configuration of GMM-UBM and i-vector systems are attributes which are mostly differentiated in their value by enumerations e. g., the development set (DevSet), the features type (VAD+39 MFCCs), and the chosen score normalisation type (scoreNormType). Public methods like *enrol, verify, evaluate* are the interface for applications using the framework, where the abstract methods *extract_features, enrol, create_probe, compare, norm* depend on the techniques of pattern recognition, meaning that these methods need to be implemented by an e. g. inherited i-vector system class which can distribute implementations in an OOP style to e. g., common libraries and toolkits.



System

- devSet : enum
- featureType : enum
- ubm_type : enum
- ubm_mixtures : uint
- scoreNormType : enum
- fusion_parameters : float*
- effectPrior : float
- enrolment_db : string
- mode : bool
- name : string*
- subsystems : System*
- subsysNames : string*

- init(devSet : enum, featureType : enum, scoreNormType : enum,
  effectPrior : float) : void
+ enrol(identity : string, sampleFiles : string*) : void
- compare(reference : string, sampleFiles : string*) : float*
+ verify(reference : string, sampleFiles : string*) : bool*
+ evaluate(scenario : enum, effectPrior: float) : float*
- *extract_features(sample : string) : float***
- *enrol(reference : string, features : float**) : void*
- *create_probe(features : float**) : float**
- *compare(templates : float**, probes : float**, mode : bool) : float***
- *norm(references : string*, probes : float**,scores : float**,
  mode : bool) : float***

Figure 36: Abstract system class diagram

Further, performance evaluations need to comprise parallel processing in order to reduce total run times. Thereby, evaluation processes can be apportioned, so that same computations are not performed twice, such as feature extraction and enrolment processes. Score normalisations can be performed single processed, because score normalisations are performable by matrix operations comprising multiple verification attempts[4] `(j)`. Fig. 37 illustrates these processes by a sequence diagram of the evaluation method from the *System* class which preserves the same evaluation procedures among different verification systems, hence evaluations are well-defined `(f)`. A scenario enumeration class is mentioned for the purposes of separating evaluation scenarios. Further, this class delivers scenario-depending enrolment data (identities and samples) and verification samples that should be compared as probes with the enrolled templates `(f)`. Multiprocessor computers give the opportunity to use parallel-processing loops in order to achieve overall evaluation in real-time where systems, e. g. based on i-vectors, may have fast scoring methods, such as matrix multiplications of template and probe i-vectors.

In research terms, evaluations on different development and evaluation (scenarios) sets might be interesting in order to obtain knowledge about e. g., scenarios of different context or text- and language-independent concurrent speaker verifications. Hence, processing databases need to be separated from development sets and evaluation scenarios. Thereby, cross-evaluations examining development sets comprising English speech applied on German-speech verification scenarios can be easily performed if the HDF5 datasets are separately stored by folder structures that are denoted according to development and scenario sets `(k)`. This concept provides the OOP processing of samples placed in different databases, such that according sample and meta-data conversions can be performed by preserving the well-defined manner of the framework `(f, g, k)`. Thus, all requirements `(a-k)` can be fulfilled by the explained OOP framework design.

5.2.3 *Data organisation*

A suitable data organisation can be designed by separating framework resources from evaluations and applications, hence by separating development from evaluation data e. g., through the folder structure. The framework's development data can further be grouped by *toolkits* (a generic hierarchy is denoted by <*TOOLKITS*>) and the framework implementation data in terms of classes, experiment setups, math and plotting (visualisation) libraries, and implemented systems such as GMM-UBM, i-vector, and a *Dummy* system (classes are marked by the @ symbol). The evaluation data can be structured by e. g., development set (DevSet) meta information (e. g. sample labels and paths), extracted features, created application and subject mod-

[4] As a speed-up illustration: multiplying two 1000-element arrays of 500-dimensional vectors by two for-loops takes 2.3575s while the matrix multiplication takes 0.0392s (≈ 98%-gain) in Matlab on a Dell Vostro 3550 having an i5 CPU@2.3GHz.

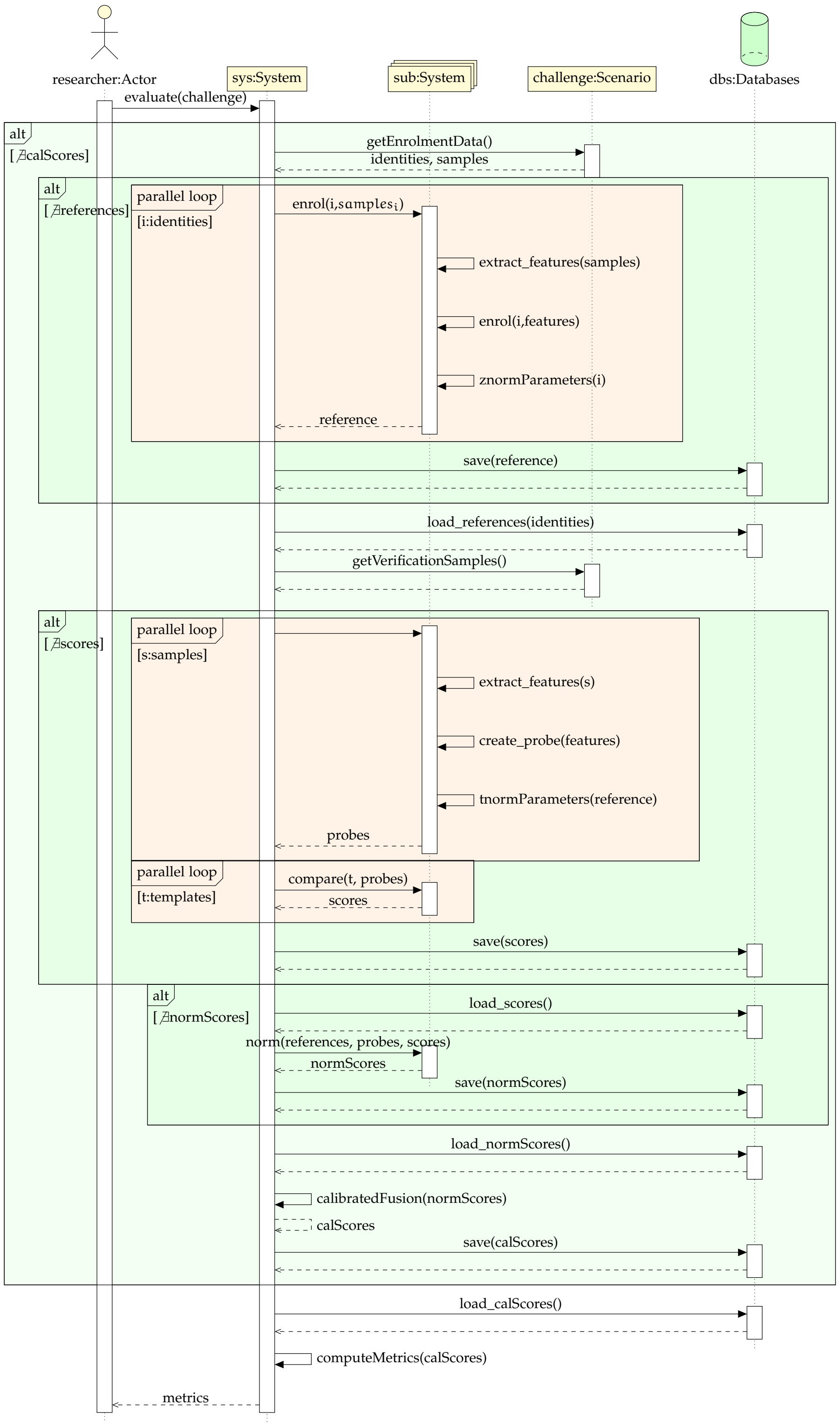
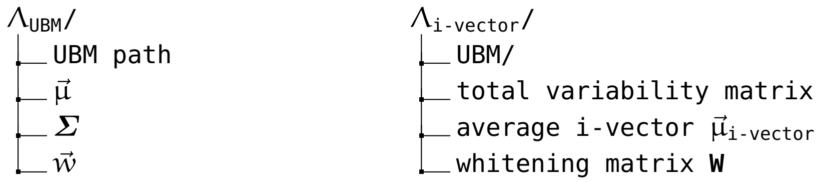


Figure 37: Evaluation processing sequence diagram

els, corpora samples, scenario meta information, and scores of system set-ups on scenarios. Further, meta information on development sets, scenarios, and scores should be provided in readable and in a binary format where **.devSet*, **.enrol*, **.verify*, and **.result* shall denote human-readable CSV formats, and **.hdf5* binary databases comprising the information by hierarchy database structures providing fast data serialisations. Thereby, model data of e. g., UBMs or i-vector hyper-parameter sets need to be separated by the underlying development data and feature types and by the UBM configurations, e. g. number of components, which are generically denoted by Λ_{UBM} i-vectors hyper-parameter sets can be denoted by $\Lambda_{\text{i-vector}}$, respectively.

The hierarchy HDF5 database structures of UBM and i-vector hyper-parameter sets are shown in fig. 38. Where the essential parameters for UBM and i-vector implementations are stored in an according binary format by e. g., means $\vec{\mu}$ and covariances $\boldsymbol{\Sigma}$.



(a) UBM

(b) i-vector

Figure 38: HDF5 database structure for hyper-parameter sets

The enrolment databases can be stored according to a system's configuration where HDF5 databases can be divided by feature types, model development sets, and e. g., UBM and i-vector parameters. Each differently configured system stores its data separately distributed over the file system: configuration changes determine the path of an enrolment database and concluding different signal processing parameters are loaded from other according HDF5 databases such as the UBM serialisation databases. For the purpose of storing references of subsystems as well, the hierarchy of the enrolment database needs to comprise this too e. g., by an *<IDENTITY>_<SUBSYSTEM>* sub-hierarchy for storing templates and normalisation parameters that belong to an identity, see fig. 39. An applicable folder structure for a speaker verification framework is presented in fig. 40.

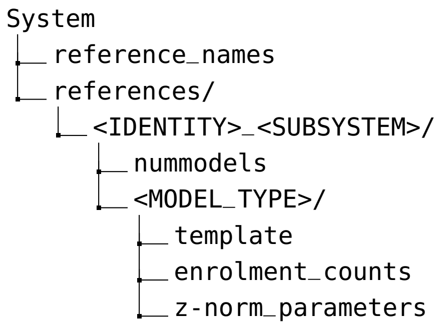


Figure 39: HDF5 structure of enrolment databases

```
kosi-ivector-framework/
├── development/
│   ├── <TOOLKITS>/
│   └── framework/
│       ├── classes/
│       ├── experiments/
│       │   └── <SETUP>.m
│       ├── math/
│       ├── plotting/
│       └── systems/
│           ├── @Dummy/
│           ├── @GMM_UBM/
│           └── @IVector/
└── evaluation/
    ├── development_sets/
    │   ├── <DevSet>.devSet
    │   └── <DevSet>.hdf5
    ├── features/
    │   └── <TYPE>/
    │       ├── <DevSet>/
    │       └── <SCENARIO>/
    ├── models/
    │   ├── <DevSet>_<TYPE>/
    │   │   ├── UBM/
    │   │   │   └── <Λ_UBM>.hdf5
    │   │   └── i-vector/
    │   │       └── <Λ_i-vector>.hdf5
    │   └── <SYSTEM>_<CONFIG>/
    │       ├── <DevSet>_<TYPE>_<UBM>.hdf5
    │       └── <DevSet>_<TYPE>_<Λ_i-vector>.hdf5
    ├── samples/
    │   └── <CORPUS>/
    ├── scenarios/
    │   ├── <SCENARIO>.enrol
    │   ├── <SCENARIO>.verify
    │   └── <SCENARIO>.hdf5
    └── scores/
        └── <SCENARIO>/
            ├── <DevSet>_<TYPE>_<SYSTEM>_<UBM>_<INFO>.result
            └── <DevSet>_<TYPE>_<SYSTEM>_<UBM>_<INFO>.hdf5
```

Figure 40: Framework folder structure

5.3 MATLAB FRAMEWORK IMPLEMENTATION

The introduced design was implemented using Matlab which provides toolkits, OOP implementations, and has effort advantages on numerical computations in terms of computation speed. Speaker verification methods are strongly based on vector and matrix operations, especially in research terms comprising many verification attempts. Common toolkits of the speaker recognition community are distributed for Matlab such as the *Joint Factor Analysis (JFA) Matlab Demo* [52] and the *BOSARIS toolkit* [40]. Those are useful for GMM-UBM, JFA, and i-vector systems implementations, and score post-processing in terms of system fusions and forensic and biometric performance evaluations. Further, Matlab's cluster analysis toolkit provides the *gmdistribution class*[5] which can be used for processing GMMs. Matlab also provides methods for serialising and deserialising HDF5 databases. Further, the *BOSARIS* toolkit refers to HDF5 databases to store e. g. it's *Score* class which was used for the framework implementation as well.

Other applied tools are taken from speech processing toolkits: Hidden Markov Model Toolkit (HTK) [47] is used for MFCC extraction and UBM training[6], *rastamat* is a Matlab toolkit for extracting and processing speech signals as well and can optionally be used on e. g. Δ-MFCC computations, and the *atip* VAD tool which was kindly provided. The application domains of the toolboxes is illustrated on fig. 41 which shows the domains with respect to the biometric system components.

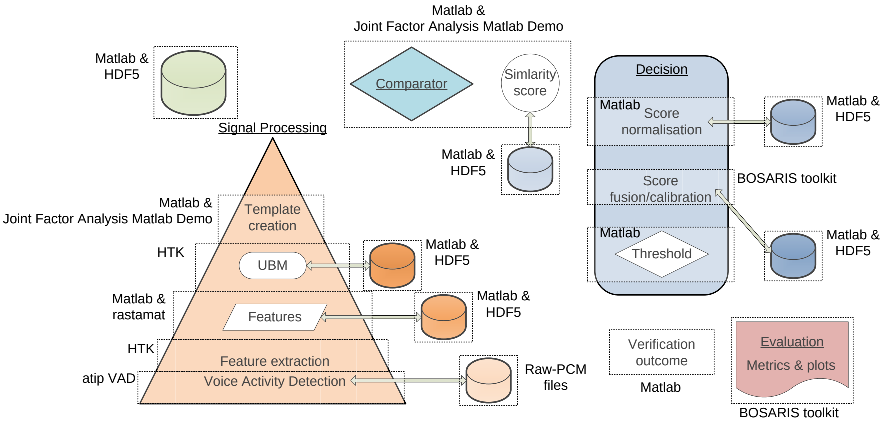


Figure 41: Application domains of Matlab toolboxes

Thereby, the total variability matrices of i-vector systems are trained by the JFA demo using its `estimate_y_and_v(...)` method which trains the JFA eigenvoice $\mathbf{V}$ matrix and speaker factors $\vec{y}$. In i-vector

[5] see: http://www.mathworks.de/de/help/stats/gmdistributionclass.html

[6] UBM training is also available within Matlab, but appeared to have computational disadvantages in terms of requested resources.

terms this equals the total variability matrix $\mathbf{T}$ and the i-vectors $\vec{i}$. Algorithm 2 briefly shows the implemented procedure of the total variability matrix training. Which differs from the eigenvoice training by the development data treatment: according to the total variability i-vector approach, see sections 3.3.4, 3.4, each sample is treated as a different source, so that the Baum-Welch statistics of all samples are stored separately. In contrast on JFA, the eigenvoices are trained by priorly grouping the Baum-Welch statistics speaker-wise. Further, JFA constrains additional channel-based supervector decomposition which is compensated by the total variability matrix by the i-vectors approach, hence the training method call contains zero values 0 and zero matrices $\mathbf{0}$ where only JFA-depending parameters are passed to the JFA demo.

```
Data: Sample set $\Omega$, sample_ids $ID_\Omega$
Result: Total variability matrix $\mathbf{T}$, ivector means $\vec{\mu}_{ivecs}$, whitening
        matrix $\mathbf{W}$
$\vec{N} \longleftarrow \emptyset$;
$\vec{F} \longleftarrow \emptyset$;
foreach $\omega$ in $\Omega$ do
    compute zero and first order Baum-Welch statistics $\vec{N}i, \vec{F}i$;
    append $\vec{N}i$ to $\vec{N}$;
    append $\vec{F}i$ to $\vec{F}$;
end
$\mathbf{T} \longleftarrow$ random();
for $n \leftarrow 1$ to $maxIter = 10$ do
    $\mathbf{T}, \vec{i} \longleftarrow$ estimate_y_and_v($\vec{F}$, $\vec{N}$, 0, $\vec{\mu}_{UBM}$, $\Sigma_{UBM}$, 0, $\mathbf{T}$, $\mathbf{0}$,
    $\mathbf{0}$, 0, $\mathbf{0}$, $ID_\Omega$);
end
$\vec{\mu}_{ivecs} \longleftarrow$ mean($\vec{i}$);
covariance eigen-decomposition $\vec{\lambda} = \mathbf{v}'\Sigma_{\vec{i}}\mathbf{v}$;
$\mathbf{W} \longleftarrow \vec{\lambda}^{-\frac{1}{2}}\mathbf{v}'$;
```

Algorithm 2: i-vector hyper-parameter set training using the JFA Matlab demo [52]

6 EVALUATION

Experiments are performed in order to answer the scientific questions about the short duration applicability of the i-vector approach, about the i-vector performance compensations on varying sample durations, and in order to verify the hypothesis on processing short duration and duration-variant samples. In this chapter the experimental set-up will be introduced first, then two evaluation datasets are described, where one comprises digit pass-phrases and the other one is the current NIST i-vector challenge corpus. The experiments are performed on a short duration and a duration variant scenario, respectively. Experiments on the short-term scenario are performed for the purpose of getting more familiar with i-vector processing, useful parameters and models.

6.1 EXPERIMENTAL SET-UP

Firstly, short duration scenario experiments show the performance of HMM-UBM, GMM-UBM and raw i-vector baseline systems. Then, emphasis is put on the effects of common i-vector processing techniques in terms of the spherical space projection as well as the effects of i-vector extraction parameters with regards to the amount of UBM components (64, 128, 256, 512, 1024, 2048), extracted i-vector factors (50, 100, 200, 300, 400, 600), and the maximum training iterations of the total variability matrix (1, 2, 5, 10) — the speaker community usually refers to 512–2048 UBM components, 400–600 i-vector factors [46, 74, 75, 76, 89], and a training of 10 iterations [100]. Further, score-level fusions of baseline and i-vector systems are evaluated, so that the amount of additional information gain by using i-vectors can be measured.

The GMM-UBM and i-vector systems are processed by the introduced framework, see chapter 5: after an energy-based VAD, 39-MFCC features are extracted using the Hidden Markov Model Toolkit (HTK), and feature warping was applied, see chapter 3, which are also used by Hegenbart [5] and Billeb [6], and the HMM-UBM system was kindly provided by atip as binary executable.

Experiments on the 2013–14 NIST i-vector challenge are performed by extending the NIST-supplied baseline system which was distributed in *Python*. Further, NIST provided 600-dimensional i-vectors without samples, but with the according sample durations. Hence, the experiment's focus is placed on the purpose of increasing the i-vector subject separability with respect to variant sample durations. Thus, the duration invariant AS-norm extension (dAS) introduced in section 4.5 was applied. Experiments on the NIST i-vector challenge are performed offline by a 5-fold cross-validation where on each run one known enrolment sample was randomly excluded from the enrolment processes and instead used for verifications, the evaluation is then reported by averaged metrics of 10 cross-validations. Further, the systems were submitted to the NIST i-vector challenge where preliminary evaluation results are presented on an online leaderboard which comprises 40% of all submitted scores where the final NIST evaluation is performed on the remaining 60% [46].

The evaluation will refer to the metrics summarised in section 2.5, in particular: EER, *FMR100*, $\mathcal{H}_{\text{norm}}^{\text{min}}$, and C_{llr} as C_{llr}^{tot}. These metrics are chosen for the purpose of appropriately measuring and coherently comparing the results among the different evaluation scenarios. Further, the algorithms are compared by the real-time factor which comprise averaged real-time measurements as $\times RT = \frac{\text{computation time}}{\text{sample duration}}$ on a *CentOS 6.3* system having an *Intel i7-3770* CPU (3.40 GHz) and 32 GB *DDR3-RAM*.

6.2 DATA DESCRIPTION

Both data sets, the atip-intern digit corpus and the NIST i-vector corpus, have differences e. g., in the sample duration (short-term and various), and in the languages where the digit corpus contains German speech and the NIST i-vectors corpus contains English speech (> 85%) as well as Spanish, Russian, Chinese, and Arabic speech[1] [46]. The following sections will describe the evaluation data used on both cross-gender all-vs.-all scenarios.

6.2.1 *Digit speech corpus*

The digit speech corpus contains the German digits zero to nine[2] spoken by more than 700 male and female subjects recorded by common telephones. Thereby, averagely 35 samples are created from each subject where two samples comprise five digits, the other 33 samples comprise three digits where the order of digits changes between subjects. All sample durations are normalised to 2.5s for three digits and 5s for five digits, respectively. Tab. 4 provides an overview on the statistical information about the digit development and scenario corpus subsets with respect to development, calibration and evaluation subsets. The development subset will be used for UBM and total variability matrix training, while the calibration subset will be used to deter-

[1] The 2013-14 NIST i-vector challenge is based upon the 2012 NIST SRE data which comprised the mentioned languages besides the German language [69, 101, 102, 103].

[2] In transcriptions of the international phonetic alphabet (IPA): /n ʊ l/ (0), /ʔ aɪ n s/ (1), /ts v aɪ/ (2), /ts v oː/ (2), /d ɾ aɪ/ (3), /f iː ɐ/ (4), /f ʏ n f/ (5), /z ɛ ç s/ (6), /z iː b n/ (7), /ʔ a x t/ (8), /n ɔ ɪ n/ (9).

mine optimal system parameters which are then validated using the evaluation subset. Hence, the baseline systems will be reported for both calibration and evaluation subsets. Where the calibration subset comprises 106 genuine and 5 830 impostor scores of 56 subjects, and the evaluation set contains 572 genuine and 171 028 impostor scores of 300 subjects. Thus, in terms of the *Rule of 3* [26] the lowest reporting error-rate being significant is approx. 1%, so that the *FMR1000* metric has no application, since it is operating at a 0.1% FMR. Hence, the biometric performance will be reported in terms of EER and *FMR100*.

Table 4: Statistics of the digit scenario's development, calibration, and evaluation sets

Set	Development	Calibration	Evaluation
Subjects	362	56	300
% female	46.9	37.5	50.6
ø samples/subject	36.0	35.0/2.0	33.1/1.9
ø sample duration	3.2s	2.5s/5.0s	2.5s/5.0s
Language	German		
Context	Digits (0-9)		

6.2.2 *2013–14 NIST i-vector challenge*

In the 2013–14 NIST i-vector challenge samples of prior NIST SREs are comprised: *the i-vectors supplied will be based on a speaker recognition system developed by the Johns Hopkins University* (JHU) *Human Language Technology Center of Excellence in conjunction with MIT Lincoln Laboratory for the 2012 NIST Speaker Recognition Evaluation (SRE)* [46], and the 2012 NIST SRE relies on data of the 2004–10 NIST SREs [69]. The joint MIT and JHU SRE'12 system description refers to 600 dimensional i-vectors that are based on a 2048-component UBM modelling 60 MFCC features (19 MFCCs and zero order MFCC as signal energy estimator with according Δ and ΔΔ), and that are extracted by *Bayesian model adaptation with an additive Gaussian noise model* [104]. VAD techniques are based on energy analyses as described in section 3.1, GMMs, and multi-frequency band analyses [104]. Further applied speech signal pre-processing techniques are described in [104]. Thus, compared to the i-vector extraction on the digit scenario, the SRE'14 i-vectors were extracted with much more efforts in compensating huge between-sample variances, i. e. phonetic content (text- and language-independence), recording devices (telephone transmitters and microphones), and environmental noise (clean, in-office, interview).

However, detailed information about the challenge i-vectors was not provided by NIST besides the according sample durations, so that comparisons to the digit corpus can only roughly be made by analysing the meta information of prior NIST SREs. Tab. 5 provides an overview according to the Linguistic Data Consortium (LDC) presentation [105] on the 2012 NIST SRE workshop and SRE’12 provided meta-information [101, 102, 103] which was analysed as far as possible; unavailable information is marked as *n/a*. Further, information is presented in subject- and reference-based (ref.) manners.

Table 5: Statistics of the NIST sets from SRE’12 and the i-vector SRE’14 challenge development and evaluation subsets

Set	SRE’12	i-vector SRE’14	
	Evaluation	Development	Evaluation
Subjects (ref.)	414 (1 818)	*n/a* (36 572)	*n/a* (1 306)
% female	60.1	*n/a*	*n/a*
ø samples/ref.	11.2/72.0	1.0	5.0/7.4
ø sample duration	212.4s/182.2s	40.8s	39.3s/39.4s
Languages (i. e.)	English, Russian, Spanish, Chinese, Arabic		
Context	Interviews & telephone calls		

In SRE’12 samples were recorded from 414 subjects of the 976 subjects of SREs’04–’10. These samples are used to create 1 818 references where on the i-vector challenge 1 306 references are enrolled. Since the current i-vector SRE’14 is not finished until 2014 April 7th, no information is available on the true amount of subjects and gender distributions during this thesis’ time frame, however a similar proportion can be assumed since the i-vector challenge strongly relies on the SRE’12. In contrast to the digit scenario, the NIST evaluation tasks differ completely: the samples comprise multi-lingual speech (besides German), the sample durations vary from less than 1s to 300s where the sample durations are log-normal distributed with an overall duration mean on 40.3s — on SRE’12 several sub-scenarios were evaluated, so that sample and duration proportions differ from the i-vector challenge. Further, the NIST SREs comprise open-content scenarios with natural speech instead of e. g., pass-phrases or digits. Hence, the NIST SREs are more difficult for robustly recognising speakers than the close-content mono-lingual digit scenario. However, in terms of enrolment sample durations the NIST SREs supply averagely more than 180s of speech where the digit scenario is bounded by 90s.

The protocols of NIST SREs [35, 46, 69, 95] aim detection costs that are computed by detection cost functions (DCFs) and that can be interpreted as normalised application-dependent entropy. The i-vector challenge refers to the minimum detection cost function (*minDCF*) that is according to the NIST protocol [46] computed by:

$$\mathit{minDCF} = \min \mathrm{DCF}(t) = \min \mathit{FNMR}(t) + 100\mathit{FMR}(t) \tag{92}$$

which accords to the minimum of the normalised Bayesian error-rate $\mathcal{H}^{\min}_{\mathrm{norm}}$ on an application-dependent prior of $\tilde{\pi} = \frac{1}{101}$, see section 2.3. In contrast, the C_{llr} metric is often referred to as well by the speaker recognition community for examining the NIST SRE tasks [25, 41, 74]. Thus, in this thesis the evaluation is also reported on the pattern matching costs in terms of $\mathcal{H}^{\min}_{\mathrm{norm}} = \mathit{minDCF}$ as the application-dependent forensic entropy and the application-independent C_{llr} metric.

6.3 short duration speaker verification

This section emphasises on short-term speaker recognition by placing focus on how applicable i-vectors are for this scenarios and which additional information the i-vector approach is able to deliver to approaches that are known to operate well on short duration scenarios e. g., HMM-UBM systems.

6.3.1 *Baseline systems*

The baseline approaches comprise HMM-UBM, GMM-UBM, and raw i-vector systems where *raw* denotes that no further processing is applied after the total variability factor analysis. Tab. 6 and fig. 42 compare system performances of systems relying on digit-based HMMs (11 per subject) and digit-independent GMMs having $C = \{64, 128, 256, 512, 1024, 2048\}$ components. Thereby, the performances are shown for both, the calibration (cal.) and the evaluation (eva.) subsets, where overall experiments on the calibration subset reached better performances, though the according data is statistically less representative according to e. g., the *Rule of 3* [29]: since only 56 subjects are used on comparisons, the approximated lowest significant error rate is 5.4% while on the evaluation subset comprising 300 subjects the lowest significant error rate would be ≈ 1.0%. However, system performance relations observed on the calibration subset accord to the observed relations on the evaluation subset: strictly proper on EER and $\mathcal{H}^{\min}_{\mathrm{norm}}$, and roughly on *FMR100* and C_{llr}. Thus, the calibration subset will be used for examining effects on i-vector-based systems in order to determine well-configured systems whose performance will then be validated on the evaluation subset.

On GMM-UBM, systems having 256- up to 1024-component UBMs reached on the evaluation subset EERs lower than 3%, *FMR100* rates lower than 5% and minimum entropies below 0.5 where UBMs having 64 and 128 components seem to result in under-fitting (EER: 4.6%,

Table 6: Baseline system performances: HMM-UBM and GMM-UBM with respect to UBM components C

Baseline	Metrics	EER	*FMR100*	$\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$	C_{llr}
HMM	Cal. set	0.3%	0.0%	0.055	0.273
	Eval. set	**0.9%**	**0.8%**	**0.172**	**0.282**
GMM $C = 64$	Cal. set	3.9%	21.6%	0.679	0.600
	Eval. set	4.6%	22.4%	0.720	0.593
GMM $C = 128$	Cal. set	2.5%	6.8%	0.506	0.500
	Eval. set	3.4%	10.7%	0.618	0.504
GMM $C = 256$	Cal. set	0.9%	0.7%	0.262	0.451
	Eval. set	2.5%	4.8%	0.422	0.462
GMM $C = 512$	Cal. set	1.2%	1.4%	0.211	0.428
	Eval. set	2.3%	3.8%	0.297	0.439
GMM $C = 1024$	Cal. set	0.5%	0.0%	0.134	0.400
	Eval. set	1.1%	1.2%	0.240	0.411
GMM $C = 2048$	Cal. set	4.1%	34.1%	0.876	0.497
	Eval. set	5.4%	37.8%	0.824	0.530

3.4%; *FMR100*: 22.4%, 10.7%; $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$: 0.720, 0.618), and 2048 components seem to result in over-fitting in terms of the short duration verification scenario (EER: 5.4%; *FMR100*: 37.8%; $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$: 0.824). The best observed GMM-UBM baseline system has 1024 components yielding 1.1% EER, 1.2% *FMR100*, 0.240 $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$ and 0.411 C_{llr}.

However, the HMM-UBM system outperforms all GMM systems on all performance metrics by yielding: 0.9% EER, 0.8% *FMR100*, 0.172 $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$ and 0.282 C_{llr}, which accords to relative achievements towards the 1024-component GMM-UBM system as: 18%, 33%, 28% and 31%, respectively. Fig. 42 presents the differences in biometric performance by the sysmtem DET curves with respect to the *Rule of 30* [26] boundaries of 30 FNMs and FMs ($\approx 5\%, \approx 0\%$).

Tab. 7 gives an overview on the baseline (raw) i-vector systems measured on the calibration subset having 400 factors and the total variability matrix was trained by 10 iterations, where the UBM component amounts are compared according to the GMM-UBM systems. On GMM-UBM systems, 1024 components yielded the lowest error and cost rates (3.7% EER, 18.8% *FMR100*, 0.430 $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$, 0.934 C_{llr}). Similarly, the 1024-component i-vector system obtained also the lowest error-rates among the raw i-vector systems, however the most raw i-vector systems perform similar to or worse than under/over-fitted GMM-UBM systems, e. g. EERs greater than 10%.

Further, all C_{llr} metrics reach a default recogniser's performance of $C_{llr} = 1$ which is caused by the cosine comparison score that is

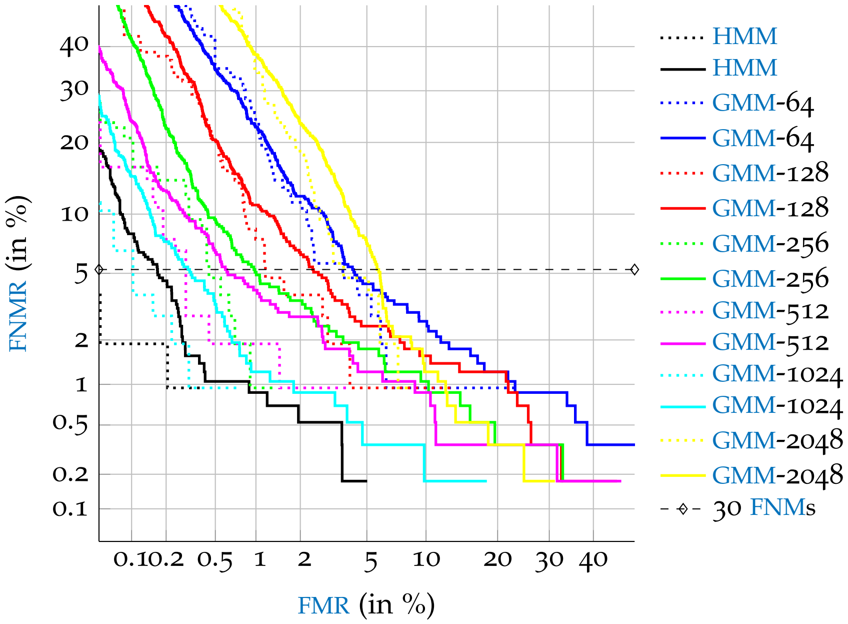


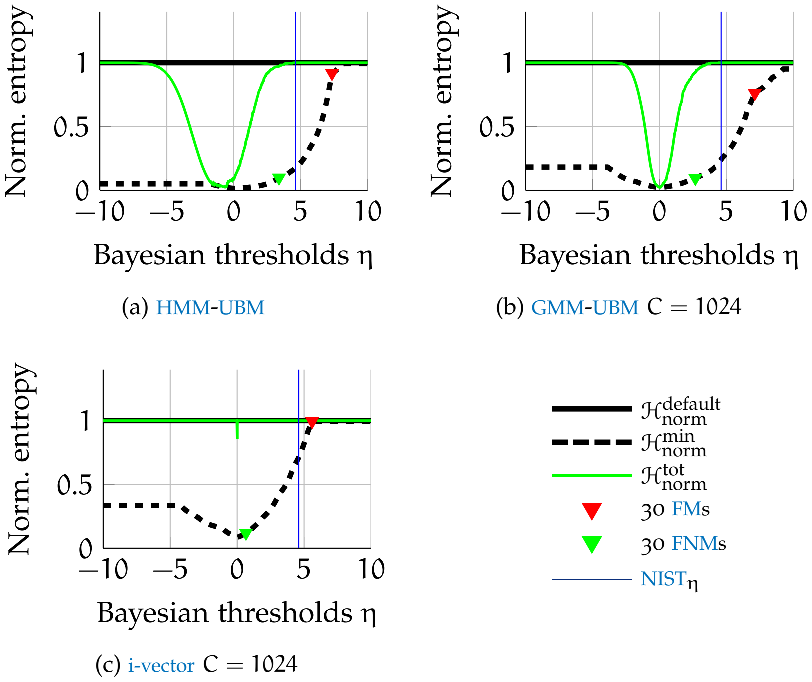
Figure 42: DET comparison of HMM-UBM and GMM-UBM systems on calibration (dashed) and evaluation (solid) subsets

Table 7: Baseline i-vector performances with respect to UBM components C on 400 factors and 10 training iterations of total variability model on cal. set

UBM C	EER	*FMR100*	$\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$	C_{llr}
$C = 64$	20.0%	65.0%	0.909	0.953
$C = 128$	17.5%	63.9%	0.834	0.948
$C = 256$	13.6%	46.6%	0.764	0.946
$C = 512$	10.1%	35.2%	0.694	0.943
$C = 1024$	**3.7%**	**18.8%**	**0.430**	**0.934**
$C = 2048$	15.8%	42.3%	0.743	0.960

bounded to the ranges of $[-1, +1]$ where the C_{llr} evaluates Bayesian thresholds in the ranges of $)-\infty, +\infty($. Fig. 43 illustrates the entropies of the 1024-UBM i-vector and GMM-UBM systems as well as of the HMM-UBM system: the GMM-UBM and HMM-UBM systems are better calibrated in terms of C_{llr} than the i-vector system of which the comparator is not based on a Bayesian reference-probe comparison.

This effect can be visualised by the normalised entropy among a feasible range of operating points $\eta \in [-10, +10]$ as shown in fig. 43. The HMM-UBM and GMM-UBM systems are well-calibrated on a few operating points close to $\eta = 0$ where the raw i-vector-1024 system also has a small range around $\eta = 0$ which is better calibrated than on all other operating points of the i-vector system, but is still too far from being well-calibrated as the huge gap between the total and the minimum entropy indicates. However, on the $\mathcal{H}_{\mathtt{norm}}^{\mathtt{min}}$ operating point $\eta = \text{NIST}_\eta$ none of the baseline systems is well-calibrated. Further, approximately until the operating point $\eta = 5$ the HMM-UBM causes the lowest entropy where on $\eta > 5$ the GMM-UBM system seems to gain advantages. In the following, effects on i-vector-based systems will be examined.



(a) HMM-UBM

(b) GMM-UBM $C = 1024$

(c) i-vector $C = 1024$

Figure 43: Normalised entropy of three baseline systems: i-vector, GMM-UBM with the same 1024-component UBM, and HMM-UBM

6.3.2 *Spherical space projection*

On duration-variant verification scenarios such as the NIST SREs, the spherical space projection was reported to yield important gains [7, 45, 82], hence effects of the spherical space projection steps (centering, whitening and length-normalisation) are analysed. In tab. 8 the results are summarised in terms of EERs by comparing single centering and whitening steps with the complete spherical space projection — the length-normalisation is already included by the cosine score computation, see eq. 82. The i-vector relocation (centering) has low effects on 64–512 UBM components, meanwhile relative EER gains of 14% and 59% were observed on 1024- and 2048-component UBMs. Thus, the projection of i-vectors into a same-origin space has more effects on higher dimensional i-vectors than on high-dimensional[3] i-vectors, hence

[3] The difference of high and higher dimensional spaces is intentioned as an order of magnitude between $C \times D = 64 \times 39 = 2\,496$ and $C \times D = 2048 \times 39 = 79\,872$.

the non-projected i-vectors of higher dimensions contain more sparse spaces than the high-dimensional i-vectors.

Table 8: Effects of i-vector spherical space projection steps on EER in % with respect to UBM components C on 400 factors and 10 training iterations of total variability model on cal. set

UBM C	Raw	Centered	Whitening-only	Spherical space
$C = 64$	20.0	19.4	35.2	6.1
$C = 128$	17.5	18.1	34.7	2.1
$C = 256$	13.6	13.8	33.3	**1.8**
$C = 512$	10.1	9.4	33.4	2.4
$C = 1024$	**3.7**	**3.2**	**31.4**	1.9
$C = 2048$	15.8	6.4	34.8	8.0

By applying only the whitening transformation which normalises variabilities and decorrelates i-vector elements, between-subject discriminant information dimishes, such that EERs of 31–35% can be observed: discriminant subject information is represented by subject-depending mean values which comprise the characteristic average factors of the average subject sub-space offset with respect to an UBM cluster. Thus, the i-vector centering needs to be considered as an important processing step to preserve characteristic information before this information is normalised and decorrelated.

However, by projecting raw i-vectors into a spherical space, huge performance gains can be observed among all i-vector systems, as all EERs are lower than 10%. Where the best EERs were observed on 256- and 1024-component UBMs with 1.8% and 1.9%, respectively. Fig. 44 compares the EERs of the spherical space transformed i-vector systems to a 5% EER cut-off threshold where systems having EERs greater than 5% shall not be further analysed within this evaluation, as 64- and 2048-component i-vector systems. Compared to the GMM-UBM baseline systems, 64-/2048-component UBMs seem not to be adequate on this verification scenario. All further experiments rely on UBMs having 128–1024 components and on spherical space projected i-vectors.

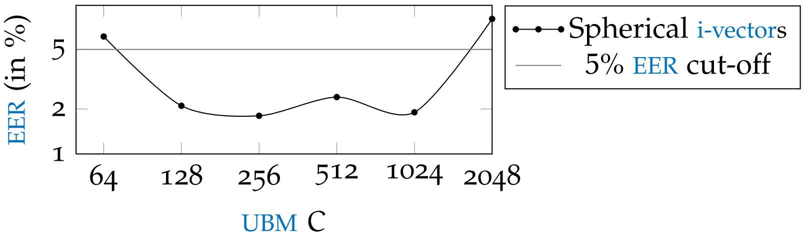


Figure 44: EERs of spherical spaces i-vectors with cut-off at 5% EER

6.3.3 *Effects of i-vector extraction parameters*

For the purpose of analysing the effects of i-vector extraction parameters, the previous fix denoted parameters of characteristic factors (400) and total variability matrix training iterations (10) are evaluated on the calibration subset. Where the i-vector dimension is varied among 50, 100, 200, 300, 400, and 600 subject factors, and the iteration amount is varied among 1, 2, 5, and 10 iterations.

Fig. 45 compares the EER performances of different factors among varying UBM components amounts: the 50-factor systems have EERs greater than 2% where the vast majority of all other factor-amount systems comprise EERs lower than 2%. The lowest EERs (1.0%-1.2%) were observed on systems comprising 200 till 600 factors. Hence, as the parameter region of interest to investigate on is bounded by 200 and 600 factors.

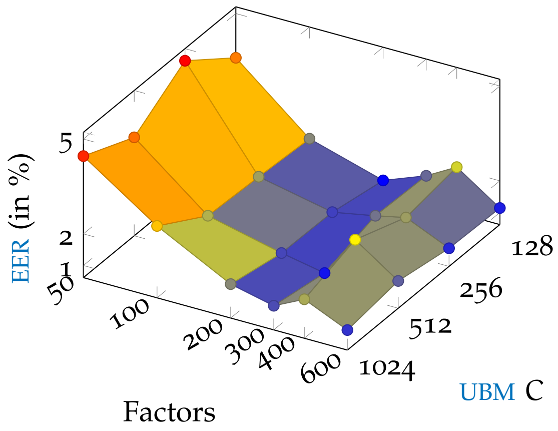
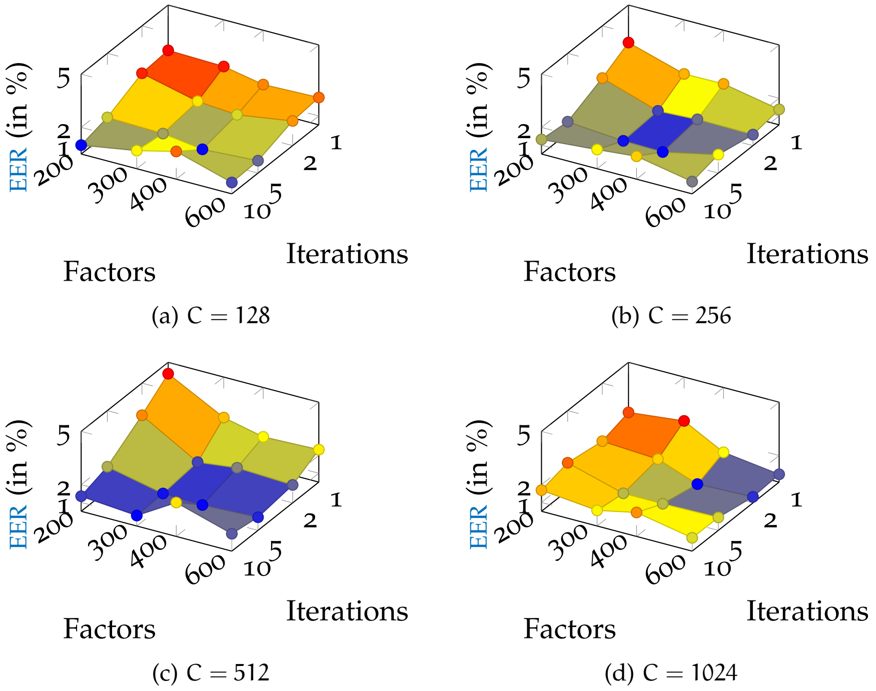


Figure 45: Performance of spherical i-vectors by different UBM amount of components C and varying factors at 10 training iterations of total variability matrix

The effect of the factor and iteration parameters is illustrated in fig. 46 where the fig. 46a–46d show the performance impacts with respect to the UBM component amounts $C = 128, 256, 512, 1024$ in terms of the EER. Among all system configurations a low amount of total variability matrix training iterations and low i-vector dimensions result in poor performances compared to more adapted and higher-dimensioned system configurations. Further, 128-component systems seem to deliver only on a few well-chosen configurations with comparable low EERs, while 256- and 1024-component systems seem to have adequate parameter regions which promise to be more robust towards evaluation data shifts. The 512-component systems seem to deliver a larger region for setting up system configurations. Though, lower EERs on 1024-component systems are observed on variability training iterations of 1 and 2 which seems that under-fitted variability models are preferred among more sparse subject representations, while on 512-component systems the intended results of sufficient variability model adaptations (2–10) and appropriate i-vector dimensions (200–600) can be observed.



(a) C = 128

(b) C = 256

(c) C = 512

(d) C = 1024

Figure 46: Effects of i-vector factor amounts and total variability matrix training iterations with respect to UBM component amounts C

Tab. 9 comprises the 16 lowest EERs measured on the calibration subset by the UBM components (128, 256, 512, 1024), the factors (200, 300, 400, 600), and the according number of variability model adaptations. As priorly observed in fig. 44 and fig. 45, the lowest EERs seem to be yielded on 256- and 1024-component UBMs (0.86%,0.73%), while on the 128-/512-component UBMs EERs of $\approx 1.00\%$ and $\approx 1.09\%$ were observed. However, for each number of UBM components the approximated best configurations are selected (bold marked) for validating them on the evaluation subset. Here 400 factors seem to be an appropriate i-vector dimension among the 128–1024 UBM component amounts.

Since on the 128- and 512-component systems two approximately similar configurations were observed, both of each are validated, so that in tab. 10 six system configurations are compared with respect to their EER, *FMR100*, $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$, C_{llr} performances on the calibration and the evaluation subsets. Over all parameter configurations, the performance shift among calibration and evaluation subsets is large as it was also observed on the baseline systems, compare tab. 6 and fig. 42. In tab. 10, the top-3 evaluation metric values for each metric are marked boldly, such that at most one value per metric is highlighted on one UBM component group. The top-3 i-vector system con-

Table 9: Effects of i-vector parameters: excerpt of best iterations in terms of EER on the calibration subset

UBM C	Factors	Iter.	EER	UBM C	Factors	Iter.	EER
128	**200**	**10**	**1.00%**	512	200	10	1.35%
	300	5	1.31%		**300**	**10**	**1.09%**
	400	**5**	**1.01%**		**400**	**5**	**1.10%**
	600	10	1.14%		600	5	1.24%
256	200	5	1.15%	1024	200	10	1.69%
	300	5	0.90%		300	5	1.17%
	400	**5**	**0.86%**		**400**	**2**	**0.73%**
	600	2	1.12%		600	2	0.84%

figurations on this short duration scenario preserve three different UBM component amounts, the system configurations and according performances in particular: 128-components, 400 factors and 5 iterations (2.1% EER, 5.1% *FMR100*, 0.347 $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$); 256-components, 400 factors, 5 iterations (2.5% EER, 5.9% *FMR100*, 0.359 $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$); and 512-components, 300 factors, 10 iterations (2.3% EER, 6.2% *FMR100*, 0.393 $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$). Hence, the 1024-UBM i-vector system was outperformed by the other component amounts, where the largest gap between calibration and evaluation subsets is observed on the 1024-component UBM, too, e. g. 0.7% to 3.2% EER and 0.1% to 9.5% *FMR100*.

Table 10: Effects of i-vector parameters: comparison calibration and evaluation subsets

UBM C	Factors	Iter.	Set	EER	*FMR100*	$\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$	C_{llr}
128	200	10	cal.	1.0%	1.0%	0.200	0.890
			eva.	2.4%	5.8%	0.373	**0.888**
	400	**5**	cal.	1.0%	1.0%	0.170	0.909
			eva.	**2.1%**	**5.1%**	**0.347**	0.911
256	**400**	**5**	cal.	0.9%	0.8%	0.211	0.909
			eva.	**2.5%**	**5.9%**	**0.359**	**0.910**
512	**300**	**10**	cal.	1.1%	1.3%	0.300	0.899
			eva.	**2.3%**	**6.2%**	**0.393**	**0.898**
	400	5	cal.	1.1%	1.3%	0.277	0.909
			eva.	2.5%	6.5%	0.403	0.909
1024	400	2	cal.	0.7%	0.1%	0.266	0.916
			eva.	3.2%	9.5%	0.452	0.916

The biometric performance of the top-3 observed i-vector configurations are visualised in fig. 47 with respect to the calibration and the evaluation subsets. While the 128-UBM i-vector system significantly outperformed both other i-vector systems on the calibration subset, all systems yield roughly the same biometric performance on the evaluation subset, where the 512-UBM system has the highest error-rates and the 128-/256-UBM systems have in the region of 0.2–1.0% FMR approximately the same error-rates. However, by taking the *Rule of 3* and *Rule of 30* boundaries [26] into account only the FNMRs/FMRs slightly greater than 1% are significant. Thus, in terms of the *FMR100* 400-dimensional i-vectors of an 128-components UBM yield the best evaluated performance among the i-vector systems without further score processing on this short-duration scenario.

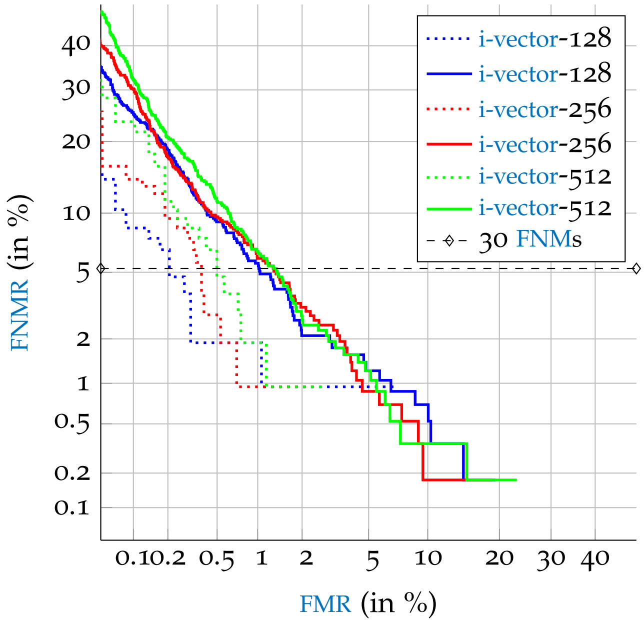


Figure 47: DET comparison of the top3-EER i-vector system configurations of 128, 256, 512 UBMs on calibration (dashed) and evaluation (solid) subsets

First experiments on the intelligent feature selection (IFS) for UBM creation, see section 3.2, did not confirm additional robustness gains, neither for GMM-UBM nor for i-vector systems in performance terms (9%–97% relative-losses in EERs). In short duration and close-context scenarios UBMs need to preserve all information, by denoting a 5%-quantile threshold for feature vector distances, UBMs seem to be modelled too sparse for i-vector-based speaker comparisons. Comparison analysis, see section 2.4, of non-IFS and IFS UBMs considering the development data showed that both UBM creations result in about the same average entropy, but in BIC terms the IFS UBMs outperformed the non-IFS UBMs by a 54% relative-gain. Hence, IFS UBMs may be considered as more robust on open-context scenarios where on close-context scenarios the acoustical space might be insufficiently modelled when features are taken out before UBM training. Thus, the presented systems comprising non-IFS UBMs are preferred for the the following research steps.

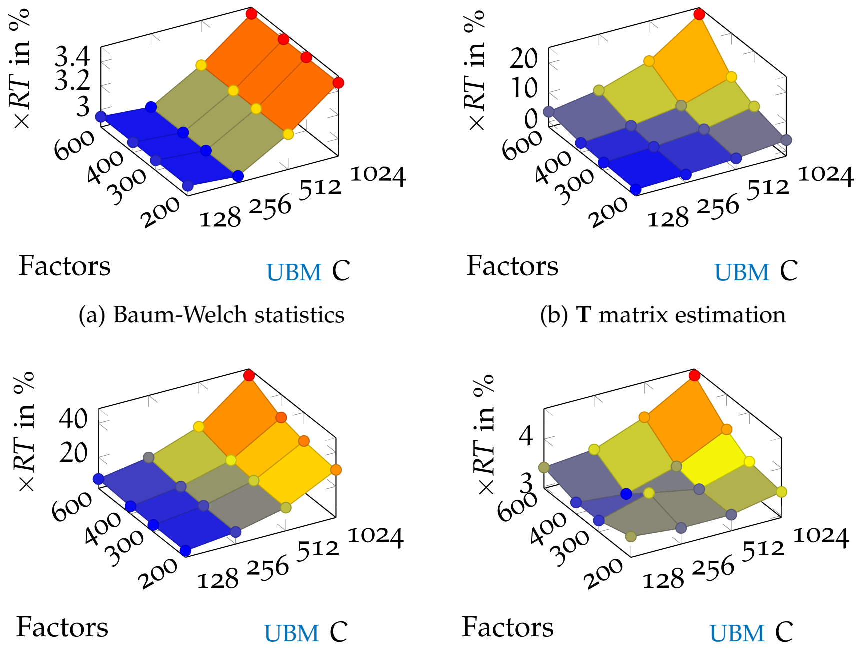
i-vector systems are able to process speech data in real-time. Fig. 48 compares efforts for Baum-Welch statistic estimation, total variability **T** matrix estimation, (re-) enrolments and verifications by the real-time factor $\times RT$. The computation time for estimating Baum-Welch statistics linearly depends on the number of UBM components C, such that if an UBM has twice as much components as another, the computation effort doubles as well. Computation efforts of estimating the **T** matrix grow exponential with respect to both the amount of i-vector factors and UBM components. Similar effects can be observed on enrolment and verification processes. Where the enrolment and re-enrolment implementations were applied that comparability to according processes on the given HMM-UBM system were consistently preserved. Thus, the computational efforts for enrolments are an order magnitude higher than the computational efforts for verifications. The highest real-time effort on verification was observed at 4.5% for an i-vector system comprising 600 factors on a 1024-component UBM[4]. Thereby, the computational main part is the estimation of the Baum-Welch statistics (3.5% $\times RT$) as well as on all other i-vector system configurations, where other i-vector processing comprise fast matrix and vector computations such as the i-vector extraction itself and the spherical space projection as described in section 3.4.

6.3.4 *Comparison of systems, system calibrations and fusions*

In comparison to the HMM-UBM and GMM-UBM baseline systems, the i-vector systems perform on enrolment and on verification processes in real-time: at most 16.2% and 3.3% $\times RT$, respectively. Where GMM-UBM systems comprise more than ten times of the computational effort on enrolments and on verifications than the real-time performances of i-vector systems relying on according UBMs which results due to the different comparator designs[5]: on GMM-UBM systems Baum-Welch statistics need to be computed twice, one time on the reference GMM and another time on the UBM, while i-vector systems only use the UBM Baum-Welch statistics to fast extract i-vectors which can be compared by the dot product. The real-time performance of the given HMM-UBM system can be located for enrolment and verification processes between the GMM-UBM systems of 256- and 512-component UBMs.

[4] Where a real-time performance peak on a 300-factor 256-component i-vector system may be the result from averaging computation durations of parallel processed verification attempts, such that this peak may be considered as an outlier.

[5] GMM-UBM scores are computed by Matlab's *gmdistribution* class, while the i-vector Baum-Welch statistics are computed by the JFA Matlab demo.



(a) Baum-Welch statistics

(b) **T** matrix estimation

(c) Enrolments re-enrolments

(d) Verification

Figure 48: Computation effort comparison of i-vector processing in terms of real-time performance

However, the i-vector systems yield the lowest computational efforts in terms of real-time processing time.

For the purpose of examining the calibration of speaker verification systems, scores of the calibration subset were used as described in section 3.4.3 to train linear score transformation matrices which are applied for system calibrations and score-level fusions on the evaluation subset scores. Tab. 12 compares the calibrated HMM-UBM, GMM-UBM, and i-vector systems by EER, *FMR100*, $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$ and C_{llr} where calibration was applied on the calibration subset scores (themselves) and on the evaluation subset scores in order to demonstrate the expected best case and evaluation C_{llr} values among the systems. Due to calibration, the HMM-UBM systems entropy could be reduced from $C_{llr} = 0.282$ to $C_{llr} = 0.054$, further on the GMM-1024 from 0.411 to 0.060, and on the i-vector systems from 0.888, 0.910, 0.898 to 0.107, 0.103, 0.101, respectively. On the short duration scenario, the baseline HMM-UBM and GMM-UBM systems perform better for all evaluation metrics than the i-vector systems: e. g. in terms of $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$, the i-vector systems yield costs between 0.347 and 0.393 where the 2014 NIST i-vector baseline system is reported to have a 0.386 $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$ on a more complex evaluation scenario (e. g., duration-variant, multi-lingual, natural language) and the HMM-/GMM-UBM systems yielded costs of 0.172 and 0.240, respectively.

The HMM models are very close-constrained towards the phonetic content and hence, more robust pattern matches can be expected

Table 11: Real-time comparison of the baseline HMM-/GMM-UBM and i-vector systems

System	Enrolment ($\approx 33 \times 2.5$s/subject)	Verification (5s/attempt)
HMM	206.2%	5.5%
GMM-128	61.6%	4.2%
GMM-256	123.3%	5.2%
GMM-512	243.2%	7.2%
GMM-1024	500.2%	10.9%
i-vector-128/400	6.1%	3.2%
i-vector-256/400	9.9%	3.1%
i-vector-512/300	16.2%	3.3%

on same-constituted speech data, while the i-vector systems are more designed for estimating characteristic subject subspaces in text- and language-independent scenarios rather than modelling e. g., phrases or digits. Further, i-vector and GMM-UBM references were created by a large enrolment collection (more than 30 samples/subject with more than 75s of speech), but on verification the GMM-UBM approach estimates to which model the probe data fits most. In contrast on the i-vector approach, template-alike probe i-vectors are extracted which are then compared to higher-sufficiently estimated template i-vectors in terms of the sample durations, e. g. $> 75s$ summed duration of all template samples compared to 5s probe sample duration. This duration-mismatch seems to increase noise on the i-vector approach. In comparison to the baseline raw i-vector systems (EERs 3.7%–20.0% and C_{llr} 0.934–0.960), the examined i-vector processing and configurations yield appropriate and calibrated performances (e. g. EERs 2.1%–2.5% and C_{llr} <0.101–0.107).

Thus, the hypothesis on short-term scenarios, see section 4.1, is confirmed: by applying appropriate modelling techniques on close-context scenarios, which are created for the purpose only to fit to the close-context patterns, more verification robustness can be obtained than by applying approaches which aim to perform robust among various phonetic contents. However, the question whether the i-vector approach is applicable on short duration verification scenarios is positively confirmed: the three resulting calibrated i-vector systems are applicable, because the observed error rates are lower than the preliminary set bounds : EER$< 5\%$ (2.1–2.5%), $FMR100 < 10\%$ (5.1–6.2%), $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}} < 0.554$ (0.347–0.393), $C_{llr} < 0.333$ (0.101–0.107), and $\times RT < 1.05 \times RT_{\text{HMM}} \approx 5.78\%$ (3.1–3.3%).

For the purpose of examining the question whether i-vector systems contribute new information towards the baseline systems, fusions of

Table 12: System calibrations of HMM-UBM, GMM-UBM, and i-vector systems

System	Calibrated	EER	*FMR100*	$\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$	C_{llr}
HMM	cal.	0.3%	0.0%	0.055	0.014
	eva.	**0.9%**	**0.8%**	**0.172**	**0.054**
GMM-1024	cal.	0.5%	0.0%	0.134	0.025
	eva.	1.1%	1.2 %	0.240	0.060
i-vector-128	cal.	1.0%	1.0%	0.170	0.056
	eva.	2.1%	5.1%	0.347	0.107
i-vector-256	cal.	0.9%	0.8%	0.211	0.049
	eva.	2.5%	5.9%	0.359	0.103
i-vector-512	cal.	1.1%	1.3%	0.300	0.057
	eva.	2.3%	6.2%	0.393	0.101

i-vector systems and the HMM-UBM system using the calibration subset are analysed where the 1024-component GMM-UBM system is not taken into account due to its exhaustively computational efforts on enrolments and verifications which are significantly outperformed by the HMM-UBM and i-vector real-time performances. Tab. 13 compares the performance of i-vector-fused systems and the fusion of the HMM-UBM system with one of each of the i-vector systems. All fusions are performed by linear logistic regression using the calibration subset for applying calibrations and fusions on the evaluation subset, see section 3.4.3.

Table 13: System fusions of i-vector and HMM-UBM systems

Systems	EER	*FMR100*	$\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$	C_{llr}
i-vector-128+256	1.9%	3.3%	0.310	0.082
i-vector-128+512	1.6%	3.6%	0.312	0.080
i-vector-256+512	1.7%	3.8%	0.321	0.077
i-vector-128+256+512	1.7%	3.2%	0.295	0.075
HMM+i-vector-128	**0.4%**	**0.1%**	**0.108**	0.439
HMM+i-vector-256	**0.4%**	**0.1%**	0.110	**0.028**
HMM+i-vector-512	**0.4%**	**0.1%**	0.110	**0.028**

By fusing i-vector systems the verification accuracy can be increased in terms of all reported evaluation metrics, such that in terms of *FMR100*, $\mathcal{H}^{\mathtt{min}}_{\mathtt{norm}}$ and C_{llr} the fusion of all three i-vector systems, denoted as *i-vector+128+256+512*, yields the lowest error-rates, costs and entropy (3.2%, 0.295, 0.075) on an 1.7% EER as an i-vector-only system, which accords to relative-gains of 32% compared with the best observed i-vector system on the calibration subset which has a 256-component UBM. Further, all fusions of single i-vector systems with the HMM-UBM system yield the best observed performances on the evaluation subset in terms of EER, *FMR100* and $\mathcal{H}^{\mathrm{min}}_{\mathrm{norm}}$. Fig. 49 illustrates the biometric performances of the HMM-UBM system, the i-vector-256 system, and the system fusions between the three i-vector systems and of the HMM-UBM system with the i-vector-256 system, denoted as *HMM+i-vector-256*.

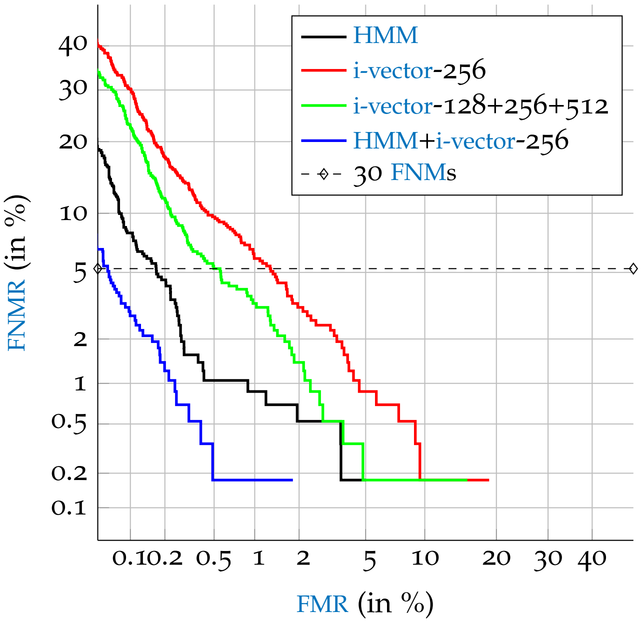
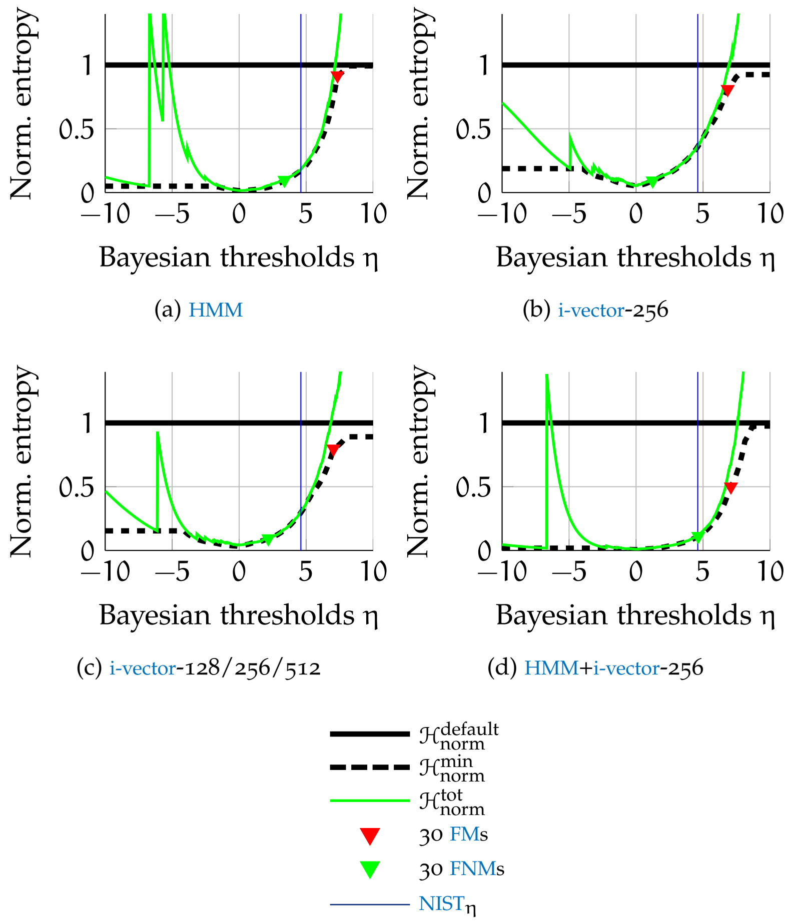


Figure 49: DET comparison of HMM-UBM, i-vector-256 systems and fusions

In terms of C_{llr}, the fusion systems comprising i-vector UBMs with more than 128 components also yield significant gains by entropies about $C_{llr} = 0.028$ which outperforms the HMM-UBM having $C_{llr} = 0.054$. Thus, i-vector systems having sufficiently fitted UBMs are able to augment information to algorithms that are well-fitted on this short-term scenario, namely the HMM-UBM approach, by a 48% rate: i-vector-256 and i-vector-512 systems contribute new information to the HMM-UBM baseline system[6].

Fig. 50 compares the normalised entropies among Bayesian thresholds within a range of $\eta \in [-10, +10]$ according to the DET-compared systems in fig. 49. All systems are well-calibrated on the evaluation operating point $\tilde{\pi} = \frac{1}{101}$ ($\mathrm{NIST}_\eta \approx 4.6$). Further, all systems obtain the lowest entropies on a $\eta = 0$ representing the intentioned EER-operating point of equal FM and FNM costs. However, linear score calibration is able to cause entropies to be greater than the default entropy on operating points with large distances to the calibration's operating point: e.g. on $\eta \approx -5$ all systems have entropy outliers

[6] Both fused systems (256/512) differ only slightly within the metrics e.g., on EERs: 0.38% vs. 0.41%.



(a) HMM

(b) i-vector-256

(c) i-vector-128/256/512

(d) HMM+i-vector-256

Figure 50: Entropy evaluation

as well as on $\eta \gg 5$, though these operating points also are outside of the significant application range which is bounded by the Rule of 30. According to the 30 FMs and FNMs boundaries, the HMM+i-vector-256 fused system's operating point approximately is within the range where both genuine and impostor scores are significantly distributed. Thereby, the HMM+i-vector-256 fused system has the lowest distance between both boundaries, such that genuine and impostor scores are separated more than within the other systems, because until the 30 FNMs boundary only 30 genuine scores are distributed among the impostor scores, which causes only small entropy and provides high biometric performances. Further, the total and minimum entropies are lower within this region than on the other systems, meaning that there are less entropy penalties due to e. g., weighted error-rate sums such as the $\mathcal{H}_{\mathtt{norm}}^{\mathtt{min}}$ or likewise *minDCF* described in eq. 92, hence there also are less scores causing mismatches that are profoundly penalised e. g., by a factor of 100 on FMs.

Comparing all four systems on all application operating-points, the HMM-UBM and i-vector-256 system have application-depending pros and cons where the i-vector-256 system gains advantages on operating points of $\eta \gg +5$ which can be increased by a fusion of all three i-vector systems. The most robust system can be established by a fusion of the HMM-UBM and the i-vector-256 systems which is a single-sample and multi-algorithm fusion. Further, according to tab. 13 the i-vector-512 system can be considered as a reliable and robust biometric verification system on this short duration scenario as well.

6.4 COMPENSATING DURATION MISMATCHES

In contrast to short duration scenarios, this section examines a duration-variant scenario which was provided by the 2013–14 NIST i-vector challenge. Within the speaker recognition community the i-vector approach was shown to perform robust on long duration samples while on short duration samples performance break-downs were observed, see section 4.3. This section aims to find out if these performance mismatches are compensable on the score-level domain. Therefore, the provided development set is analysed first, and afterwards the in section 4.5 proposed duration-invariant AS-norm extension is evaluated on the NIST challenge which also aims to validate the second hypothesis of this thesis which is proposed in section 4.4.

6.4.1 *Data analysis*

NIST provided 600 dimensional i-vectors, and their according sample durations which are log-normal distributed as shown in fig. 51. Most of the development sample durations are in the 20–40 second range, i. e. these samples are influencing development set based i-vector processing such as the spherical space projection.

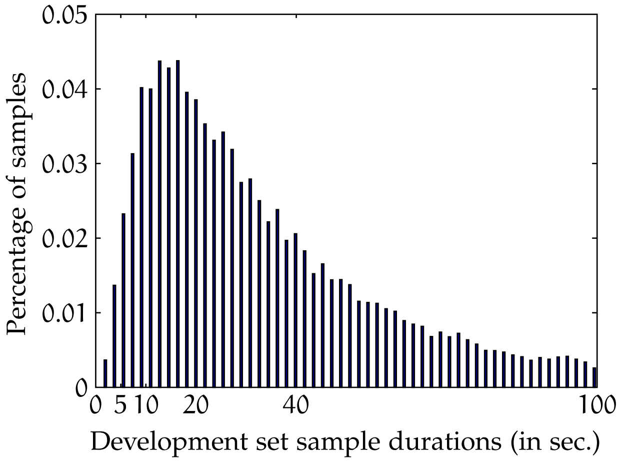


Figure 51: Log-normal distributed sample durations of development set data with respect to i-vector sufficiency classes

The vast majority of development samples are located in the quality classes of Λ_{full} (34.7%), Λ_{40} (31.1%), and Λ_{20} (23.1%), then: Λ_{10} (9.0%), and Λ_5 (2.1%), for specifications of the duration classes see section 4.5. Intentionally, all i-vectors have been centralised to the origin by mean-subtraction during the preparation of the baseline system, but an unpaired Student t-test of independence showed that i-vector elements have significantly different mean-values compared between all development set i-vectors and with respect to each sufficiency class. Tab. 14 compares the amount of significantly independent i-vector elements according to their sample durations assuming equal variance[7]. Once the spherical space projection has been applied, Λ_{40} i-vectors exhibit the lowest significant offset to the space origin by having the second most impact on both i-vector processing due to their representative amount. Further, the most-sufficient i-vectors have the greatest gap compared to all development set i-vectors and to each sufficiency class with at least 140/600 significant different mean positions. Hence, within the subspace of Λ_{40} i-vectors seem to be between the subspaces of high-insufficient and high-sufficient i-vectors. An opposite effect could be observed on short-duration samples, where the according i-vectors have larger mean-differences to i-vectors of more than 20 seconds than to i-vectors of comparable short duration samples (less than 20 seconds). This effect may be caused due to high variability of insufficient estimated i-vectors of short-duration samples, i. e. i-vectors of less than 20 second samples are distributed in subspaces that are closer to themselves than to more-sufficiently estimated i-vectors.

Table 14: Student t-test of independent i-vector elements with respect to sufficiency classes

Development set	all	Λ_{full}	Λ_{40}	Λ_{20}	Λ_{10}
Λ_5	84	141	91	66	44
Λ_{10}	142	230	140	70	
Λ_{20}	132	246	118		
Λ_{40}	35	180			
Λ_{full}	172				

That is, offset vectors can be assumed for each sufficiency group, which effects the cosine score values due to angle changes between i-vectors[8]. These facts underline the need for compensating scoring statistics with respect to sample durations. Hence, first investigations were performed by applying Fisher's LDA, see section 2.4, in order determine the most characteristic factors among the development

[7] Results of an unpaired Student t-test assuming unequal variances yielded negligible differences in the results.

[8] Which actually is additive noise that should be well-compensable by, e. g. G-PLDA scoring [25, 88].

i-vectors. However, first (insignificant) gains could be yielded by applying the LDA transformation according to the probe sample duration. Generalised dimension reduction experiments increased the error-rates instead of lowering them, hence duration-dependant treatments seem to be necessary. Generalised dimension reduction experiments increased the error-rates instead of lowering them, hence duration-dependant treatments seem to be necessary. Thus, this thesis research emphasis is placed on the duration invariant AS-norm (dAS) for the purpose of compensating i-vector subspace mismatches due to acoustic holes as described in section 4.5.

6.4.2 *Duration invariant AS-norm*

Focusing on the baseline system the highest performance loss in entropy terms of $\mathcal{H}_{\text{norm}}^{\text{min}}$ is observed for low-durational samples, see tab. 15. As can be seen, Λ_5 i-vectors yielded the highest entropy with 0.932 which is very close to a random recognisers performance of $\mathcal{H}_{\text{norm}}^{\text{default}} = 1$. i-vectors stemming from the class with the longest sample duration yielded the best observed $\mathcal{H}_{\text{norm}}^{\text{min}}$, i. e. 0.219.

Table 15: Duration group performances: avg. $\mathcal{H}_{\text{norm}}^{\text{min}}$

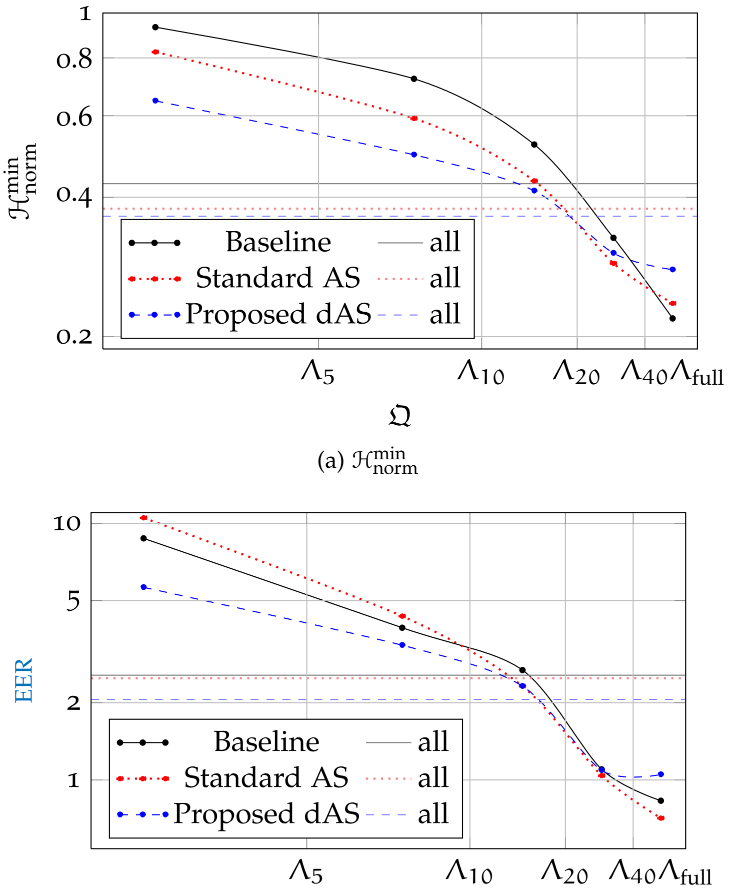
System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.932	0.721	0.520	0.327	**0.219**
AS-norm	0.824	0.592	0.434	**0.288**	0.236
dAS-norm	**0.646**	**0.494**	**0.413**	0.303	0.279

However, on all other quality classes of insufficient i-vectors both AS-normalisations yield significant gains where the dAS-norm outperforms both other systems on samples shorter than 20 seconds. On 20–40 second samples both normalisations could outperform the baseline approach, where AS-norm without duration-sensitive extension achieved the best $\mathcal{H}_{\text{norm}}^{\text{min}}$ for Λ_{40} i-vectors. Hence, AS-norm is necessary on insufficiently estimated i-vectors, and the proposed duration-based extension can yield up to 19.1% more relative-gain than the standard AS-norm. In terms of biometric recognition performance both AS-norm approaches outperform the baseline as well, see tab. 16.

Table 16: Duration group performances: avg. EER

System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	8.74	3.92	2.68	1.10	0.83
AS-norm	10.51	4.35	**2.32**	**1.04**	**0.71**
dAS-norm	**5.63**	**3.35**	**2.32**	1.09	1.05

Again, the proposed dAS-norm yields significant gains on samples shorter than 20 seconds on which a performance break-down for the standard AS-norm can be observed. However, on higher-sufficient i-vectors the standard AS-norm outperforms both other systems, which could be caused by the non-duration-invariance of the Λ_{full} i-vectors. EER and $\mathcal{H}_{\text{norm}}^{\text{min}}$ performance comparisons among quality classes $\mathfrak{Q}$ are shown in fig. 52.



(a) $\mathcal{H}_{\text{norm}}^{\text{min}}$

(b) EER

Figure 52: Biometric performance of i-vector sufficency classas

Across the entire set of classes the proposed duration-based AS-norm outperforms both other systems, see tab. 17. In summary, the proposed dAS-norm yields a 19.5% relative-gain in EER, a 32.6% relative-gain in *FMR100*, and a 15.0% relative-gain in $\mathcal{H}_{\text{norm}}^{\text{min}}$ compared to the baseline system on the cross-validation. Further, the dAS-norm significantly outperforms the standard AS-norm which can also be seen in fig. 53.

The results were approved by the preliminary evaluation of the 2013–14 NIST i-vector challenge, where the application of the standard AS-norm resulted in a 14.2% relative-gain, and the duration-invariant extension resulted in a 19.2% relative-gain in $\mathcal{H}_{\text{norm}}^{\text{min}}$.

Table 17: System performances: avg. EER, *FMR100*, $\mathcal{H}^{\text{min}}_{\text{norm}}$

System	EER	*FMR100*	$\mathcal{H}^{\text{min}}_{\text{norm}}$	Challenge[9]
Baseline	2.56	5.15	0.428	0.386
AS-norm	2.49	4.48	0.378	0.331
dAS-norm	**2.06**	**3.47**	**0.364**	**0.312**

Fig. 53 compares the best cross-validation systems according to the minimum entropy within a DET diagram. The dAS-norm improves the biometric performance of the baseline system at all operating points, while the standard AS-norm mainly yields gains in high-secure regions, i. e. operating points at low FMRs. In this regions both AS-normalisations exhibit equal recognition accuracy.

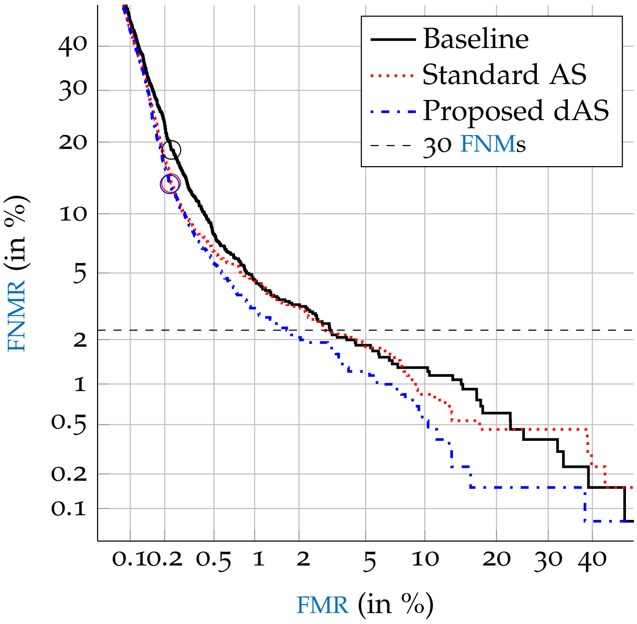


Figure 53: Systems DET: best systems from 10 cross-validations according to their minimum entropy.

Hence, the proposed duration-invariant AS-norm extension is applicable to a larger range of scenarios compared to the standard AS-norm. While the dAS-norm only obtains slightly lower error-rates on $\mathcal{H}^{\text{min}}_{\text{norm}}$-operating points compared to the standard AS-norm, more advantages of the duration-invariant treatment are observed within wide-application entropy evaluations.

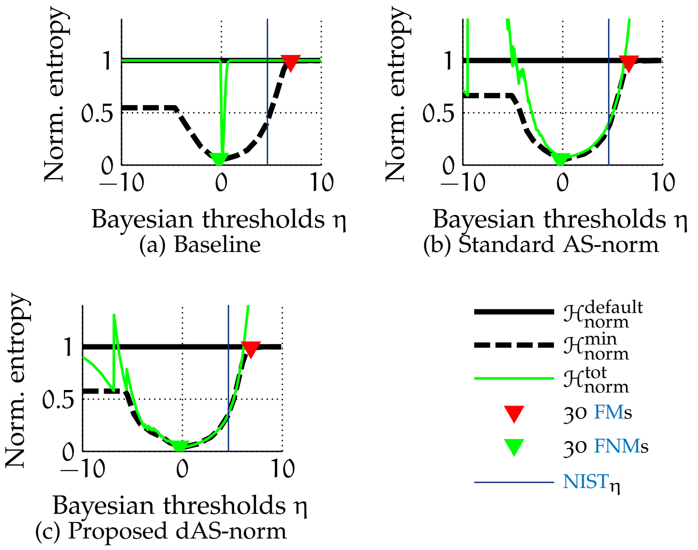
6.4.3 *Wide-application entropy analyses*

Tab. 18 compares the total C_{llr} of the three systems over all scores, and among each quality class. On Λ_5 i-vectors the baseline and the standard AS-norm perform similar to or worse than a random recogniser, and on samples having more than 5 seconds the standard AS-norm significantly outperforms the baseline system. The lowest application-independent entropy on high-sufficient i-vectors was measured for the standard AS-norm with $C_{llr} = 0.05$, representing a very low cost of the LLR-scores. However, on sample durations lower than 40 seconds the dAS-norm outperforms both other approaches by yielding a maximum LLR cost of $C_{llr} = 0.35$ on high-insufficient Λ_5 i-vectors. Overall the suggested AS-norm extension exhibits the lowest application-independent entropy by yielding relative-gains of 88.8% and 41.2%, respectively. Fig. 54 illustrates the C_{llr} gains with which emphasis is

Table 18: Average entropy comparison: all scores & duration-groups

System	all	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.89	0.95	0.93	0.92	0.89	0.86
AS-norm	0.17	1.18	0.41	0.18	0.08	**0.05**
dAS-norm	**0.10**	**0.35**	**0.20**	**0.11**	**0.07**	0.07

also placed on robustness, i. e. systems which do not require score-calibration, because total and minimum entropy are equal on well-calibrated systems. Due to the cosine scoring all scores of the baseline system lie within the range $[-1, +1]$, hence, the lowest entropy. The smallest difference between total and minimum entropy was observed on $\eta \approx 0$, on any other operating point the baseline system is effected by huge mis-calibrations. Calibration-improvements were



(a) Baseline

(b) Standard AS-norm

(c) Proposed dAS-norm

Figure 54: Entropy comparison of the NIST baseline and submitted systems

gained by the standard AS-norm, which delivers adequate calibration for different application-points. However, the suggested duration-invariant score normalisation yields well-calibrated scores on the vast majority of application-points, which have significant error-rates. That is, the proposed duration-invariant enables an enhanced statistical treatment of quality classes, which is approved by a very low overall entropy emission in terms of C_{llr}.

Hence, duration-based performance mismatches could be compensated on score-level domain by simulating i-vector subspace relocations after i-vector comparisons by taking further knowledge into account about i-vectors of similar quality shape e. g., in terms of the template and probe sample durations. Thus, the stated hypothesis on duration-invariant scenarios could be confirmed by the robustness gains in terms of biometric performances, the NIST evaluation metric *minDCF* or likewise $\mathcal{H}^{\text{min}}_{\text{norm}}$, the C_{llr} metric as the application-independent entropy, and wide-application ranged entropy analyses.

6.5 summary and discussion

The evaluations were performed on short duration and duration-invariant scenarios. Thereby, the short duration scenario comprised German digits with sample durations under 5s while the duration-invariant NIST scenario relies on multi-lingual speech data from interviews and telephone calls where sample durations vary from about 1s to 300s. Experimental investigations were performed with the purposes to find out whether the i-vector approach is applicable on short duration scenarios, if new information can be gained by using i-vector systems together with well-established approaches in short duration speaker recognition, and whether there are (simple) methods to increase the robustness of i-vector systems on duration-variant application scenarios.

Starting from HMM-UBM, GMM-UBM, and raw i-vector baseline systems, the effects of i-vector processing steps e. g., spherical space projection, and parameter configurations, e. g. amounts of UBM components and characteristic factors, were analysed on short duration scenarios. By examining proper system configurations using a calibration subset, three i-vector systems could be established and evaluated on a calibration-distinct evaluation subset where all three reported i-vector systems yielded acceptable performances, especially in terms of first implementation steps on i-vector systems. In adequate real-time comparisons to the baseline systems, the i-vector systems significantly outperformed the HMM-UBM and GMM-UBM systems on both, enrolment and verification processes. However, in this short duration scenario, where e. g. the HMM-UBM approach models the close-content data very precisely, the i-vector approach models more general observations in terms of sample-depending UBM cluster offsets from which subject-characteristic factors are extracted. Thereby, the scenario's samples are also strongly influenced by the phonetic content of ten digits which is a close-content set, thus extracted i-vectors seem to suffer from these effects while e. g. the HMM-UBM approach takes advantage of these. Hence, the hypothesis on short duration scenarios (section 4.2) could be confirmed which states that on short duration scenarios similar-content and -quality analyses, e. g. by digit-based HMMs, will achieve more robustness than more generalised comparisons as by i-vectors.

For the purpose of evaluating the i-vector information gain towards known approaches, score-level system fusions have been applied using information of the calibration subset. Thereby, a fusion of the three i-vector systems outperformed those, and further, by fusing single i-vector systems with the HMM-UBM system the best performances could be observed in terms of EER, *FMR100*, $\mathcal{H}^{\text{min}}_{\text{norm}}$, and C_{llr}. By outperforming the HMM-UBM system in terms of C_{llr} (overall entropy), the fusion systems proved that the i-vector systems deliver new information to existing systems, such that all examined metrics could be improved by 36%–88% relative-gains. The fusions of the HMM-UBM system with either i-vector systems of a 256-component UBM on 400 factors and a 512-component UBM on 300 factors obtained similar results, in particular: 0.4% EER, 0.1% *FMR100*, 0.110 $\mathcal{H}^{\text{min}}_{\text{norm}}$, and 0.028 C_{llr}. It is important to note that this sort of fusion is a single-sample and multi-algorithm fusion, where all comparison algorithms need to compensate the same noise effects of one sample and cannot take information of re-captured samples into account for achieving higher robustness.

Further, i-vectors of duration-variant scenarios were examined in order to analyse compensation techniques for quality-mismatch-based performance break-downs. Analyses on the development set data from the multi-lingual 2013–14 NIST i-vector challenge confirmed that i-vector subspaces vary depending on the sample duration as a quality metric. Hence, a score-level domain compensation approach was examined that was motivated by the well-established AS-norm and extended by the motivation of the hypothesis described in section 4.4 which states that sample comparisons need to transform features of quality-differing samples into a common space. By establishing the duration-invariant AS-norm extension significant gains to the NIST baseline system and the standard AS-norm can be observed in terms of EER, *FMR100*, $\mathcal{H}^{\text{min}}_{\text{norm}}$, and C_{llr} overall samples. On low duration samples the dAS-norm also outperforms the standard AS-norm. In contrast, the standard AS-norm seems to have advantages on higher duration samples. However, by comparing the three systems on a wide-application range, the duration compensation method, that is proposed in this thesis, outperforms the baseline and standard AS-norm systems in terms of well-calibration. Hence, additional score calibration sets and transformation matrices as applied in the short duration scenario evaluation might on duration-invariant score normalisations not be as necessary as on short duration applications.

7 CONCLUSIONS AND FUTURE PERSPECTIVES

This thesis examined speaker verification using i-vectors on short duration and duration-variant scenarios. The focus was put on analysing the behaviour of i-vector-based systems in short duration scenarios in comparison with well-known approaches and on compensating performance mismatches on duration-variant scenarios. A basic i-vector system was constructed, proven to be applicable, and shown to contribute new information towards existing speaker verification systems, such that significant performance and robustness gains were yielded. For reproducibility purposes an ISO-based verification framework design was created and implemented. Thus, the presented evaluation results are sound. Further, the robustness of i-vector systems was increased by a duration-invariant score normalisation that was proven on an international evaluation of the NIST.

Hence, i-vector systems are applicable within industry scenarios e. g., randomised pass-phrase-based verification scenarios on contact centers, and also for scenarios of unknown sample durations such as continuous speaker verification on mobile devices. Further, i-vector processing can be used on forensic applications e. g., separating speakers among samples. However, other applications of HMM- and GMM-based patter recognition such as medical investigations of the human heart rate or image segmentations might also take advantages of the i-vector approach by training the total variability matrix more with respect to other contents rather than to biometric subjects.

Future investigations may examine Bayesian scoring methods such as the Gaussian probabilistic LDA (G-PLDA) that was motivated by the face recognition community for the purpose of comparing templates and probes under the assumption of noise effects which is also well-established within the speaker recognition community. Since the recognition robustness relies on the UBM quality, other UBM types might be evaluated where dimension decoupled GMMs promise gains in terms of the BIC. Other researches may consider to apply speech signal features which lately became very popular among the speaker recognition community[1] for e. g. feature-level fusions such as i-vector concatenations. Further, for the purpose of providing reproducibility among all biometric fields in research and industry terms, an open-source framework extending the proposed speaker verification framework will be implemented in a free usable language which aims at maintainability and fast computations as well.

[1] I. e.: perceptual linear prediction (PLP) features, mean Hilbert envelope coefficients (MHECs), power normalised cepstral coefficients (PNCCs), phone LLRs (PLLRs), and prosodic polynomial contours (ProsPols).

Part III

APPENDIX

A NOTATIONS

The speaker recognition community uses many terms describing the same objects. Further, there are conflicts with other communities like the standardised Biometric community. In the following the notations used in this thesis are described and relations to the ISO-harmonised biometric vocabulary [21], and to the community notation are referred to as well.

General notations

Notation	Description	Source
a	Scalar	
$\vec{a}, \vec{A}$	Vector	
$\mathbf{A}$	Matrix	
N	Number of	
n	Control variable	
R	Rank	
μ	Mean	
$\sigma^2, \boldsymbol{\Sigma}$	Variance, covariance matrix	
P	A-posteriori probability	
p	A-priori probability	
ld	Logarithmus dualis ld, or binary logarithm lb	
$\mathcal{H}$	Entropy	
χ	Subject/speaker (specific)	

Data sets

Notation	Description	Source
DevSet	Development set [46]: data set used to create an application database [21]	
EnrolSet	Enrolment set: data set used to create an enrolment database [21] by using speaker-specific meta information and speaker samples	
VerifySet	Verification set: data set used to measure the biometric performance [21]	

Data sets

Notation	Description	Source
Φ/ϕ	Scenario set: combination of enrolment and verification sets	

Normalisation sets

Notation	Description	Source
$\mathfrak{Z}$	Data set of system default impostors to simulate attacks on a speakers reference: model-specific parameters of impostor score distributions can be determined and used for later on normalisation [12, 106]	
$\mathfrak{T}$	Data set of system default impostors which is used to determine sample score distribution parameters during verifications (tests) to normalise the probe [12, 106]	

Hypotheses

Notation	Description	Source
H	Hypothesis whether the claim of an presented individual is genuine or not	
H_0	Null-hypothesis: genuine (target trial, claimed identity equals speaker identity)	[85, 106]
H_A	Alternative hypothesis: impostor, non-target trial [95], claimed identity not equals speaker identity	

Samples & Features

Notation	Description	Source
Ω/ω	Sample set/sample	
VAD	Voice Activity Detection, speech activity detection, utterance detection	[12]
MFCC	Mel-frequency-cepstral coefficient	[12, 106]
$\boldsymbol{\Psi}/\vec{\psi}$	Feature matrix/vector	
D	Dimension of $\vec{\psi}$	[44]

Models

Notation	Description	Source
Λ	Hyper-parameter set of JFA/i-vector modell	[23]
λ	Speaker model	[12, 44]

GMM-UBM

Notation	Description	Source
GMM	Gaussian-mixture-model consisting of components of D-dimensional Gaussian distributions	[12, 44, 106]
UBM	Universal background model	[12, 44, 106]
$\vec{\mu}$	Mean supervector	
$\vec{\Sigma}$	Covariance supervector	
C/c	GMM's components	[50]
w_c	Component's weight, $\sum_{c \in C} w_c = 1$	[44]
$\mathcal{N}$	Gaussian distribution	
$P_c(\vec{\psi})$	Posteriori probability of the alignment of $\vec{\psi}$ to a component c	[44]

Baum-Welch statistics

Notation	Description	Source
$\vec{N}$	Zero order statistics, $\vec{N}_c(\vec{\psi}) = P_c(\vec{\psi})$	[23, 50, 81]
$\mathbf{F}$	First order statistics, $\mathbf{F}_c(\vec{\psi}) = P_c(\vec{\psi})\vec{\psi}$	[23, 50, 81]
$\mathbf{S}$	Second order statistics, $\mathbf{S}_c(\vec{\psi}) = diag\left(P_c(\vec{\psi})\vec{\psi}\vec{\psi}'\right)$	[23, 50, 81]
$\vec{F}(\omega)$	Centered first order statistics	[23, 50, 81]
$\vec{S}(\omega)$	Centered second order statistics	[23, 50, 81]

Joint Factor Analysis

Notation	Description	Source
JFA	Joint Factor Analysis, $\Lambda = (\vec{\mu}_{UBM}, \mathbf{V}, \mathbf{U}, \mathbf{D}, \boldsymbol{\Sigma})$, $\vec{\mu}_{\omega,\chi} = \vec{\mu}_{UBM} + \mathbf{V}\vec{y}(\chi) + \mathbf{U}\vec{x}(\omega) + \mathbf{D}\vec{z}(\lambda_\chi)$	[23, 50, 81]

Joint Factor Analysis

Notation	Description	Source
$\mathbf{V}$	Eigenvoice matrix	[23, 50, 81]
$\mathbf{U}$	Eigenchannel matrix	[23, 50, 81]
$\mathbf{D}$	Residual matrix	[23, 50, 81]
$\vec{y}(\chi)$	Speaker factors	[23, 50, 81]
$\vec{x}(\omega)$	Channel factors	[23, 50, 81]
$\vec{z}(\lambda_\chi)$	Residual factors	[23, 50, 81]

i-vector

Notation	Description	Source
i-vector	Identity-vector, $\Lambda = (\vec{\mu}_{UBM}, \mathbf{T}, \boldsymbol{\Sigma})$, $\vec{\mu}_{\omega,\chi} = \vec{\mu}_{UBM} + \mathbf{T}\vec{\imath}(\chi)$	[7]
$\mathbf{T}$	Total variability matrix	[7]
$\vec{\imath}(\chi)$	Total factors, identity-vector, denoted as w in [7]	
LDA	Linear discriminant analysis for dimension reduction (Fisher's LDA)	[7]
PLDA	Probabilistic LDA	[24]

Scores

Notation	Symbol	Source
S	Score	[31]
LLR	Log-likelihood ratio	[40, 95]
t	Threshold	[40]

Performance

Notation	Symbol	Source
$\mathfrak{E}$	Evidence measure/recognition system	[18, 40]
$\times RT$	Real-time factor	
FTC	Failure-to-capture	[26]
FTA	Failure-to-acquire $FTA = FTC + (1 - FTC)\frac{N_{\text{failed acquisitions}}}{N_{\text{acquisitions}}}$	[26]
FTE	Failure-to-enrol	[26]

Performance

Notation	Symbol	Source
FMR	False match rate, type I error, false positive rate, matched impostors	[26]
$FNMR$	False non-match rate, type II error, false negative rate, non-matched genuines	[26]
FAR	False acceptance rate, false alarm probability in [40, 95], $FAR = (1 - FTA)FMR$	[26]
FRR	False rejection rate, miss probability in [40, 95], $FRR = FTA + (1 - FTA)FNMR$	[26]
EER	Equal error rate	[31]
$HTER$	Half total error rate on a certain t e. g., $HTER = \frac{FNMR(t_{DevSet,EER})+FMR(t_{DevSet,EER})}{2}$	[107]
$FMR100$	FNMR at $FMR = 1\%$	
AUC	Area under curve	
C_{FM}	Cost of a false match, C_{fa} in [40, 95]	
C_{FNM}	Cost of a false non-match, C_{miss} in [40, 95]	
π	Genuine probability, $\operatorname{logit}\pi = \log\frac{\pi}{1-\pi}$, P_{tar} in [95]	[40]
$\tilde{\pi}$	Effective prior, $\tilde{\pi} = \frac{\pi C_{FNM}}{\pi C_{FNM}+(1-\pi)C_{FM}}$	[40]
DCF	Detection cost function, $DCF(t) = \pi C_{FNM} FNMR(t)$ $+ (1-\pi)C_{FM} FMR(t)$	[40, 95]
$\mathcal{H}_{\text{err}}$	Empirical Bayes error-rate (entropy), $\mathcal{H}_{\text{err}}(\tilde{\pi}) = \tilde{\pi} FNMR(-\operatorname{logit}\tilde{\pi})$ $+ (1-\tilde{\pi})FMR(-\operatorname{logit}\tilde{\pi})$	[40]
$\mathcal{H}_{\text{norm}}$	Normalised entropy, $\mathcal{H}_{\text{norm}} = \frac{\mathcal{H}_{\text{err}}}{\min(\tilde{\pi},1-\tilde{\pi})} = DCF_{\text{norm}}$	[40]
η	Bayesian threshold, $\eta = -\operatorname{logit}\tilde{\pi} = \log\frac{C_{FM}}{C_{FNM}} - \operatorname{logit}\pi$	[40, 95]
$actDCF$	Actual (normalised) DCF at $t = \eta$	[40, 95]
$minDCF$	Minimum (normalised) DCF, threshold-independent	[40, 95]
C_{llr}	Goodness of LLRs, by integrating out all operating points $\tilde{\pi}$ of $\mathcal{H}_{\text{err}}$	[40, 95]

B QUALITY CLASSIFICATION OF BIOMETRIC CHARACTERISTICS

Jain et al. [1] refer to seven measurement requirements to qualify whether or not a physical or biological trait can be used as a biometric characteristic:

- Universality: each person should have the characteristic,
- Distinctiveness: any two persons should be sufficiently different in terms of the characteristic,
- Permanence: the characteristic should be sufficiently invariant (with respect to the comparison criterion) over a period of time,
- Collectability: the characteristic can be measured quantitatively,
- Performance, which refers to the achievable recognition accuracy and speed, the resources required to achieve the desired recognition accuracy and speed, as well as the operational and environmental factors that affect the accuracy and speed,
- Acceptability, which indicates the extent to which people are willing to accept the use of a particular biometric identifier (characteristic) in their daily lives, and,
- Circumvention, which reflects how easily the system can be fooled using fraudulent methods.

C PROPER SCORING RULE EXAMPLES

Popular examples for proper scoring rules are [18, 37, 38]:

- Brier or Quadratic score (strictly proper):

  $S_{\text{brier}} = \sum_{i=1}^{N} p_i^2 - 1,$

  $L_{\text{brier}}(p, q) = -\sum_{i=1}^{N} (\delta_{qi} - p_i)^2 = 2p_q - \sum_{i=1}^{N} {p_i}^2 - 1;$

- Spherical score (strictly proper):

  $$S_{\text{sph}} = \left(\sum_{i=1}^{N} p_i^{\alpha}\right)^{\alpha^{-1}},$$

  $$L_{\text{sph}}(p, q) = \frac{p_q^{\alpha-1}}{(\sum_{i=1}^{N} p_i^{\alpha})^{\frac{\alpha-1}{\alpha}}};$$

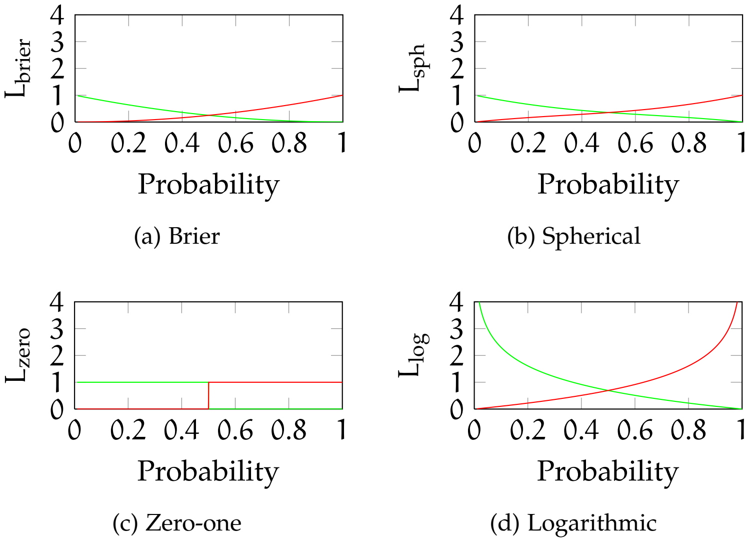
- Zero-one loss (not-strictly proper):

  $$L_{\text{zero}}(p, q) = \begin{cases} 0, \text{if } q = p, \\ 1, otherwise \end{cases};$$

- Logarithmic score as negative Shannon entropy (strictly proper):

  $S_{\log}(p) = -\mathcal{H}_p = \sum_{i=1}^{N} p_i \log p_i,$

  $L_{\log}(p, q) = \log p_q.$



(a) Brier

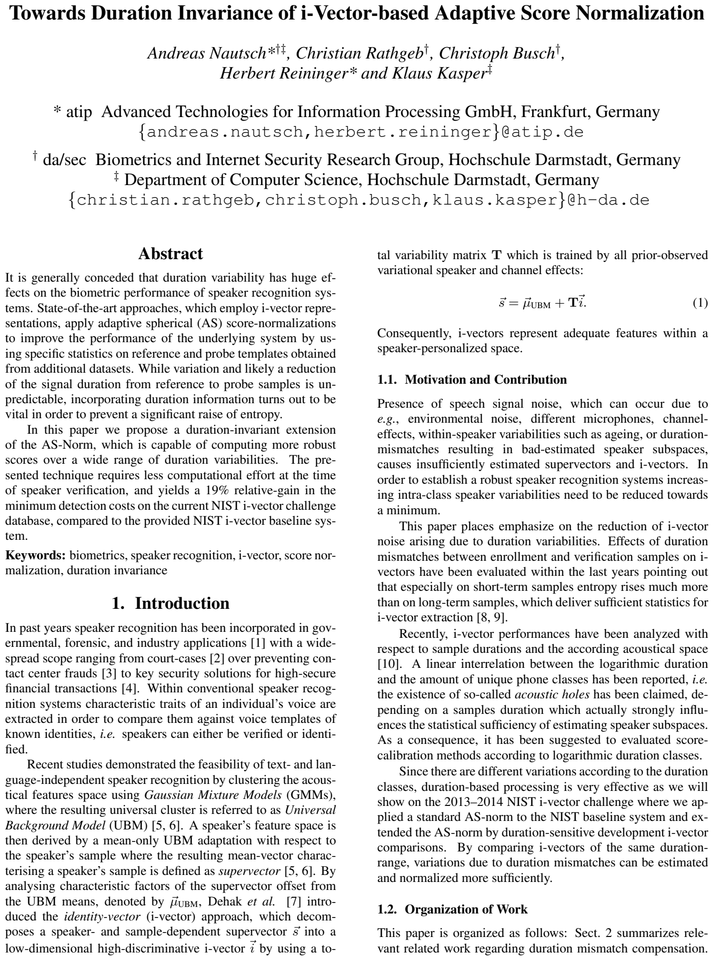
(b) Spherical

(c) Zero-one

(d) Logarithmic

Loss functions on genuine (green) and impostor (red) detection errors

D SUBMITTED SPEAKER ODYSSEY 2014 PAPER



Towards Duration Invariance of i-Vector-based Adaptive Score Normalization

Andreas Nautsch†‡, Christian Rathgeb†, Christoph Busch†,*
Herbert Reininger and Klaus Kasper‡*

* atip Advanced Technologies for Information Processing GmbH, Frankfurt, Germany
{andreas.nautsch,herbert.reininger}@atip.de

† da/sec Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
‡ Department of Computer Science, Hochschule Darmstadt, Germany
{christian.rathgeb,christoph.busch,klaus.kasper}@h-da.de

Abstract

It is generally conceded that duration variability has huge effects on the biometric performance of speaker recognition systems. State-of-the-art approaches, which employ i-vector representations, apply adaptive spherical (AS) score-normalizations to improve the performance of the underlying system by using specific statistics on reference and probe templates obtained from additional datasets. While variation and likely a reduction of the signal duration from reference to probe samples is unpredictable, incorporating duration information turns out to be vital in order to prevent a significant raise of entropy.

In this paper we propose a duration-invariant extension of the AS-Norm, which is capable of computing more robust scores over a wide range of duration variabilities. The presented technique requires less computational effort at the time of speaker verification, and yields a 19% relative-gain in the minimum detection costs on the current NIST i-vector challenge database, compared to the provided NIST i-vector baseline system.

Keywords: biometrics, speaker recognition, i-vector, score normalization, duration invariance

1. Introduction

In past years speaker recognition has been incorporated in governmental, forensic, and industry applications [1] with a widespread scope ranging from court-cases [2] over preventing contact center frauds [3] to key security solutions for high-secure financial transactions [4]. Within conventional speaker recognition systems characteristic traits of an individual's voice are extracted in order to compare them against voice templates of known identities, *i.e.* speakers can either be verified or identified.

Recent studies demonstrated the feasibility of text- and language-independent speaker recognition by clustering the acoustical features space using *Gaussian Mixture Models* (GMMs), where the resulting universal cluster is referred to as *Universal Background Model* (UBM) [5, 6]. A speaker's feature space is then derived by a mean-only UBM adaptation with respect to the speaker's sample where the resulting mean-vector characterising a speaker's sample is defined as *supervector* [5, 6]. By analysing characteristic factors of the supervector offset from the UBM means, denoted by $\vec{\mu}_{\text{UBM}}$, Dehak *et al.* [7] introduced the *identity-vector* (i-vector) approach, which decomposes a speaker- and sample-dependent supervector $\vec{s}$ into a low-dimensional high-discriminative i-vector $\vec{i}$ by using a total variability matrix $\mathbf{T}$ which is trained by all prior-observed variational speaker and channel effects:

$$\vec{s} = \vec{\mu}_{\text{UBM}} + \mathbf{T}\vec{i}. \qquad (1)$$

Consequently, i-vectors represent adequate features within a speaker-personalized space.

1.1. Motivation and Contribution

Presence of speech signal noise, which can occur due to *e.g.*, environmental noise, different microphones, channel-effects, within-speaker variabilities such as ageing, or duration-mismatches resulting in bad-estimated speaker subspaces, causes insufficiently estimated supervectors and i-vectors. In order to establish a robust speaker recognition systems increasing intra-class speaker variabilities need to be reduced towards a minimum.

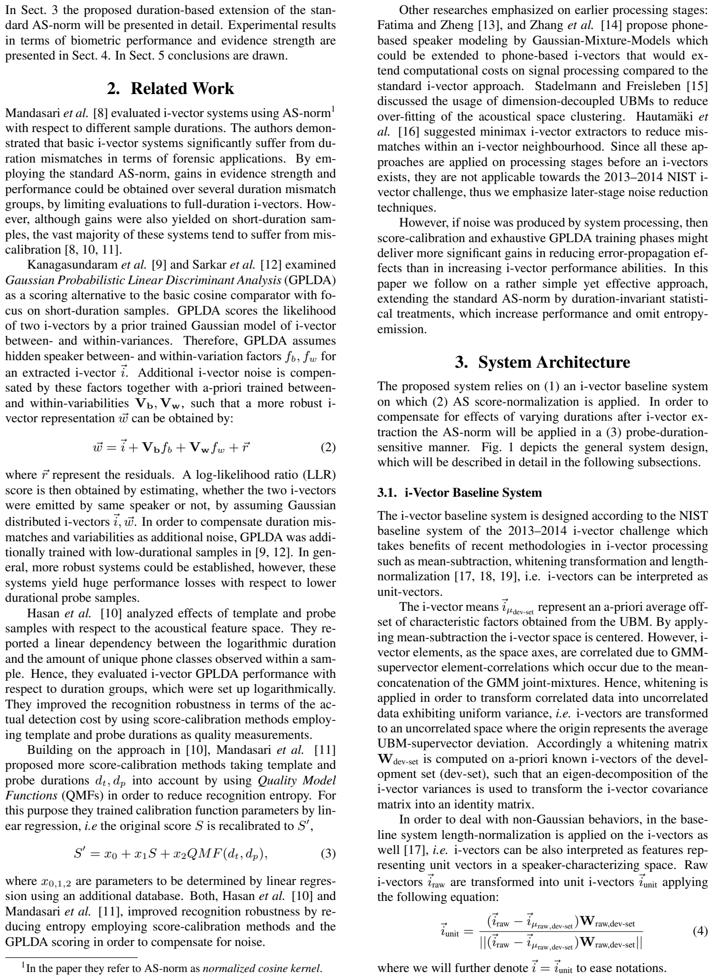
This paper places emphasize on the reduction of i-vector noise arising due to duration variabilities. Effects of duration mismatches between enrollment and verification samples on i-vectors have been evaluated within the last years pointing out that especially on short-term samples entropy rises much more than on long-term samples, which deliver sufficient statistics for i-vector extraction [8, 9].

Recently, i-vector performances have been analyzed with respect to sample durations and the according acoustical space [10]. A linear interrelation between the logarithmic duration and the amount of unique phone classes has been reported, *i.e.* the existence of so-called *acoustic holes* has been claimed, depending on a samples duration which actually strongly influences the statistical sufficiency of estimating speaker subspaces. As a consequence, it has been suggested to evaluated score-calibration methods according to logarithmic duration classes.

Since there are different variations according to the duration classes, duration-based processing is very effective as we will show on the 2013–2014 NIST i-vector challenge where we applied a standard AS-norm to the NIST baseline system and extended the AS-norm by duration-sensitive development i-vector comparisons. By comparing i-vectors of the same duration-range, variations due to duration mismatches can be estimated and normalized more sufficiently.

1.2. Organization of Work

This paper is organized as follows: Sect. 2 summarizes relevant related work regarding duration mismatch compensation.



In Sect. 3 the proposed duration-based extension of the standard AS-norm will be presented in detail. Experimental results in terms of biometric performance and evidence strength are presented in Sect. 4. In Sect. 5 conclusions are drawn.

2. Related Work

Mandasari *et al.* [8] evaluated i-vector systems using AS-norm[1] with respect to different sample durations. The authors demonstrated that basic i-vector systems significantly suffer from duration mismatches in terms of forensic applications. By employing the standard AS-norm, gains in evidence strength and performance could be obtained over several duration mismatch groups, by limiting evaluations to full-duration i-vectors. However, although gains were also yielded on short-duration samples, the vast majority of these systems tend to suffer from miscalibration [8, 10, 11].

Kanagasundaram *et al.* [9] and Sarkar *et al.* [12] examined *Gaussian Probabilistic Linear Discriminant Analysis* (GPLDA) as a scoring alternative to the basic cosine comparator with focus on short-duration samples. GPLDA scores the likelihood of two i-vectors by a prior trained Gaussian model of i-vector between- and within-variances. Therefore, GPLDA assumes hidden speaker between- and within-variation factors f_b, f_w for an extracted i-vector $\vec{i}$. Additional i-vector noise is compensated by these factors together with a-priori trained between- and within-variabilities $\mathbf{V_b}, \mathbf{V_w}$, such that a more robust i-vector representation $\vec{w}$ can be obtained by:

$$\vec{w} = \vec{i} + \mathbf{V_b} f_b + \mathbf{V_w} f_w + \vec{r} \tag{2}$$

where $\vec{r}$ represent the residuals. A log-likelihood ratio (LLR) score is then obtained by estimating, whether the two i-vectors were emitted by same speaker or not, by assuming Gaussian distributed i-vectors $\vec{i}, \vec{w}$. In order to compensate duration mismatches and variabilities as additional noise, GPLDA was additionally trained with low-durational samples in [9, 12]. In general, more robust systems could be established, however, these systems yield huge performance losses with respect to lower durational probe samples.

Hasan *et al.* [10] analyzed effects of template and probe samples with respect to the acoustical feature space. They reported a linear dependency between the logarithmic duration and the amount of unique phone classes observed within a sample. Hence, they evaluated i-vector GPLDA performance with respect to duration groups, which were set up logarithmically. They improved the recognition robustness in terms of the actual detection cost by using score-calibration methods employing template and probe durations as quality measurements.

Building on the approach in [10], Mandasari *et al.* [11] proposed more score-calibration methods taking template and probe durations d_t, d_p into account by using *Quality Model Functions* (QMFs) in order to reduce recognition entropy. For this purpose they trained calibration function parameters by linear regression, *i.e* the original score S is recalibrated to S',

$$S' = x_0 + x_1 S + x_2 QMF(d_t, d_p), \tag{3}$$

where $x_{0,1,2}$ are parameters to be determined by linear regression using an additional database. Both, Hasan *et al.* [10] and Mandasari *et al.* [11], improved recognition robustness by reducing entropy employing score-calibration methods and the GPLDA scoring in order to compensate for noise.

Other researches emphasized on earlier processing stages: Fatima and Zheng [13], and Zhang *et al.* [14] propose phone-based speaker modeling by Gaussian-Mixture-Models which could be extended to phone-based i-vectors that would extend computational costs on signal processing compared to the standard i-vector approach. Stadelmann and Freisleben [15] discussed the usage of dimension-decoupled UBMs to reduce over-fitting of the acoustical space clustering. Hautamäki *et al.* [16] suggested minimax i-vector extractors to reduce mismatches within an i-vector neighbourhood. Since all these approaches are applied on processing stages before an i-vectors exists, they are not applicable towards the 2013–2014 NIST i-vector challenge, thus we emphasize later-stage noise reduction techniques.

However, if noise was produced by system processing, then score-calibration and exhaustive GPLDA training phases might deliver more significant gains in reducing error-propagation effects than in increasing i-vector performance abilities. In this paper we follow on a rather simple yet effective approach, extending the standard AS-norm by duration-invariant statistical treatments, which increase performance and omit entropy-emission.

3. System Architecture

The proposed system relies on (1) an i-vector baseline system on which (2) AS score-normalization is applied. In order to compensate for effects of varying durations after i-vector extraction the AS-norm will be applied in a (3) probe-duration-sensitive manner. Fig. 1 depicts the general system design, which will be described in detail in the following subsections.

3.1. i-Vector Baseline System

The i-vector baseline system is designed according to the NIST baseline system of the 2013–2014 i-vector challenge which takes benefits of recent methodologies in i-vector processing such as mean-subtraction, whitening transformation and length-normalization [17, 18, 19], i.e. i-vectors can be interpreted as unit-vectors.

The i-vector means $\vec{i}_{\mu_{\text{dev-set}}}$ represent an a-priori average offset of characteristic factors obtained from the UBM. By applying mean-subtraction the i-vector space is centered. However, i-vector elements, as the space axes, are correlated due to GMM-supervector element-correlations which occur due to the mean-concatenation of the GMM joint-mixtures. Hence, whitening is applied in order to transform correlated data into uncorrelated data exhibiting uniform variance, *i.e.* i-vectors are transformed to an uncorrelated space where the origin represents the average UBM-supervector deviation. Accordingly a whitening matrix $\mathbf{W}_{\text{dev-set}}$ is computed on a-priori known i-vectors of the development set (dev-set), such that an eigen-decomposition of the i-vector variances is used to transform the i-vector covariance matrix into an identity matrix.

In order to deal with non-Gaussian behaviors, in the baseline system length-normalization is applied on the i-vectors as well [17], *i.e.* i-vectors can be also interpreted as features representing unit vectors in a speaker-characterizing space. Raw i-vectors $\vec{i}_{\text{raw}}$ are transformed into unit i-vectors $\vec{i}_{\text{unit}}$ applying the following equation:

$$\vec{i}_{\text{unit}} = \frac{(\vec{i}_{\text{raw}} - \vec{i}_{\mu_{\text{raw,dev-set}}})\mathbf{W}_{\text{raw,dev-set}}}{||(\vec{i}_{\text{raw}} - \vec{i}_{\mu_{\text{raw,dev-set}}})\mathbf{W}_{\text{raw,dev-set}}||} \tag{4}$$

where we will further denote $\vec{i} = \vec{i}_{\text{unit}}$ to ease notations.

[1] In the paper they refer to AS-norm as *normalized cosine kernel*.

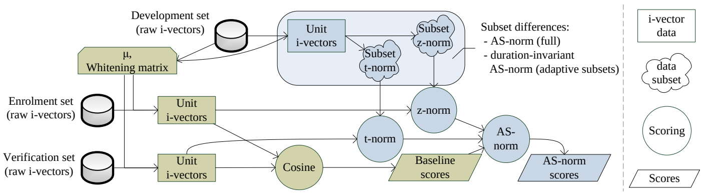


Figure 1: Basic operation mode of the proposed duration-sensitive speaker recognition system.

Speaker references are created by averaging multiple enrollment i-vectors resulting in noise-robust templates [19, 20], which can be further interpreted as a sample-concatenated simulation where higher-sufficient Baum-Welch statistics are averaged, such that more speaker-characterizing i-vectors are extracted. At the time of verification the cosine similarity comparison between template and probe i-vectors is used according to the NIST baseline system [19]:

$$S(\vec{i}_t, \vec{i}_p) = \frac{\vec{i}_t^T \ \vec{i}_p}{||\vec{i}_t|| \, ||\vec{i}_p||} \tag{5}$$

where the i-vectors are already length-normalized, *i.e.* only the numerator term of Eq. 5 is required for score computations.

3.2. AS-Norm

For the purpose of applying standard score-normalization methods by preserving the symmetry between i-vectors, Kenny [21] introduced the *spherical normalization* (s-norm). Thereby the *zero score-normalization* (z-norm) computes the score mean $\mu_{\text{z-norm}}$ and standard deviation $\sigma_{\text{z-norm}}$ of a template i-vector compared against an i-vector collection $\mathfrak{Z}$, and the *test score-normalization* (t-norm) compares similar parameters $\mu_{\text{t-norm}}, \sigma_{\text{t-norm}}$ of a probe i-vector against an i-vector collection $\mathfrak{T}$. Hence, a verification score S can be normalized by centering impostor scores having unit variance by known impostor score distributions with respect to a template i-vector and of a probe i-vector as if it was an impostor i-vector,

$$S' = \frac{1}{2}\left(\frac{S - \mu_{\text{z-norm}}}{\sigma_{\text{z-norm}}} + \frac{S - \mu_{\text{t-norm}}}{\sigma_{\text{t-norm}}}\right). \tag{6}$$

The AS-norm S' differs from s-norm by the scores which are used to compute the z/t-statistics: rather than using all scores, only the most competitive scores (*e.g.* top-100) are applied to model according speaker cohorts. Dehak *et al.* [22] applied the AS-norm on i-vectors and showed that the score normalization can be already applied on comparison-level as a normalized cosine scoring,

$$S(\vec{i}_t, \vec{i}_p) = \frac{(\vec{i}_t - \vec{i}_{\mu_{\text{z-norm}}})^T (\vec{i}_p - \vec{i}_{\mu_{\text{t-norm}}})}{||\boldsymbol{\Sigma}_{\text{z-norm}} \vec{i}_t|| \, ||\boldsymbol{\Sigma}_{\text{t-norm}} \vec{i}_p||} \tag{7}$$

where $\vec{i}_{\mu_{\text{z-norm}}}, \vec{i}_{\mu_{\text{t-norm}}}$ denote mean i-vectors of z- and t-norm sets, and $\boldsymbol{\Sigma}_{\text{z-norm}}, \boldsymbol{\Sigma}_{\text{t-norm}}$ are according diagonal covariance matrices.

3.3. Proposed Duration-invariant Approach

In order to build upon the idea of only taking significant comparisons into account, AS-norm is adapted to differentiate between probe sample durations. As previously mentioned, the presence of acoustic holes increases the entropy of shorter voice samples, which motivates the construction of different i-vector sufficiency-classes. Hence, the AS-norm is extended such that only comparisons are used for AS-parameter estimation that have the same quality as the current probe presented for verification.

In terms of duration as a quality metric, Q quality classes can be denoted as: $\mathfrak{Q} = \{\Lambda_0, \ldots, \Lambda_Q\}$ representing i-vector sufficiency classes. Samples are then associated by their logarithmic duration d_s to a sufficiency class Λ_c by the lowest log-duration distance,

$$\arg_{\Lambda_c} \min |\log(d_s) - \log(d_{\Lambda_c})|. \tag{8}$$

3.3.1. *i-vector sufficiency classes*

In the proposed system duration-based groups are defined for the sufficiency classes, where we limit the number of quality classes to $Q = 5$, *i.e.* obtained results can be directly compared to those reported in [10, 11]. It was found that evaluations carried out for the adaptive log-duration range from Eq. 8 yielded no significantly different results. Thus, sufficiency classes are denoted according to the researches on acoustic holes of Hasan *et al.* [10] and Mandasari *et al.* [11] and summarized in Table 1, where Λ_{full} is intended to comprise all expected high-sufficient i-vectors which might cause non-optimal results, but preserves low-computation efforts.

Table 1: Sufficiency classes and corresponding durations

Sufficiency class	Duration
Λ_5	0–5 sec
Λ_{10}	5–10 sec
Λ_{20}	10–20 sec
Λ_{40}	20–40 sec
Λ_{full}	$\geq$ 40 sec

3.3.2. *Parameter Estimation*

For the z- and t-norm parameter AS-cohorts are pre-selected in various ways:

- z-norm simulates impostor verifications on averaged enrollment templates, thus only $\mathfrak{Z}$ i-vectors will be used which are belong to the same sufficiency class as the probe i-vector:

$$\mathfrak{Z} = \{\vec{i}_{\Lambda_{d_p}} \,|\, \max_{\text{top100}} S(\vec{i}_t, \vec{i}_{\Lambda_{d_p}})\}, \tag{9}$$

- t-norm simulates impostor verifications comparing the probe i-vector to templates of the development set, where enrolled speakers have full i-vectors, where the vast majority of durations are higher than 60 seconds, *i.e.* only $\mathfrak{T}$ i-vectors will be used extracted from samples with longest durations:

$$\mathfrak{T} = \{\vec{i}_{\Lambda_{>60}} \,|\, \max_{\text{top100}} S(\vec{i}_t, \vec{i}_{\Lambda_{>60}})\}. \tag{10}$$

3.3.3. Score estimation and expected improvements

The proposed duration-adaptive extension of AS-norm normalizes the scores according to Eq. 6. By placing emphasis on duration-based sufficiency classes, recognitions are treated duration-invariant, *i.e.* normalized scores are expected to be distributed without creating entropy due to duration-mismatches. Further, an overall improvement can be expected, since scores of all sufficiency classes are normalized to more similar distributions of genuine and impostor scores. As a consequence, no additional entropy is expected,which could arise due to score-distribution mismatches by fixed across-classes thresholds.

Fig. 2 illustrates how duration-differing samples will be processed by the proposed AS-norm extension.

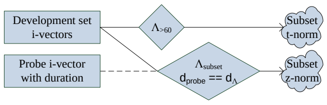

Development set i-vectors
Probe i-vector with duration
$\Lambda_{>60}$
Λ_{subset}
$d_{\text{probe}} == d_{\Lambda}$
Subset t-norm
Subset z-norm

Figure 2: Processing duration-differing samples by suggested duration-based AS-norm extension.

4. Experimental Evaluation

Experiments are carried out on the 2013–2014 NIST i-vector challenge dataset [19] in order to evaluate the baseline, the standard AS-norm, and the duration-based AS-norm extension. We performed ten 5-fold cross-validations[2] on the enrollment database, and we submitted each system also on the i-vector challenge where preliminary results were computed by NIST using 40% of the whole evaluation set.

4.1. Experimental Set-up

The NIST i-vector challenge dataset consists of $1\,306$ speaker identities within enrollment and verification sets. For each identity 5 enrollment i-vectors are given with the according sample duration. The verification set contains $9\,634$ probe i-vectors with the according sample duration as well. Further, a development set of $36\,572$ independent i-vector with sample durations is given for feature space estimations, independent of the evaluation data[3].

Focusing on performance evaluation, we place emphasize on the biometric recognition performance in terms of the Equal-Error-Rate (EER), and the false non-match rate at a 1% false match rate (FMR100). In accordance to the ISO/IEC IS 19795-1 [23] the FNMR of a biometric system defines the proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample. By analogy, the FMR defines the proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template. As score distributions overlap EERs are obtained, i.e. the system error rate where FNMR = FMR. Further, we estimate the entropy and biometric performance in terms of the application-dependent[4] minimum detection cost function [19]

$$\text{minDCF} = \min\, \text{FNMR} + 100\,\text{FMR}, \tag{11}$$

and the application-independent entropy by the log-likelihood ratio cost [25] of genuine and impostor scores SG, SI

$$C_{\text{llr}} = \frac{\sum_{g \in SG} \operatorname{ld}(1 + \frac{1}{e^{S_g}})}{2|SG|} + \frac{\sum_{i \in SI} \operatorname{ld}(1 + e^{S_i})}{2|SI|}. \tag{12}$$

4.2. Data Analysis

The provided i-vectors exhibit 600 dimensions, and their according sample durations are log-normal distributed as shown in Fig. 3. Most of the development sample durations are in the 20–40 second range, *i.e.* these samples are influencing development set based i-vector processing such as mean-subtraction and whitening.

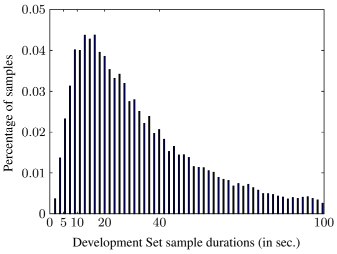


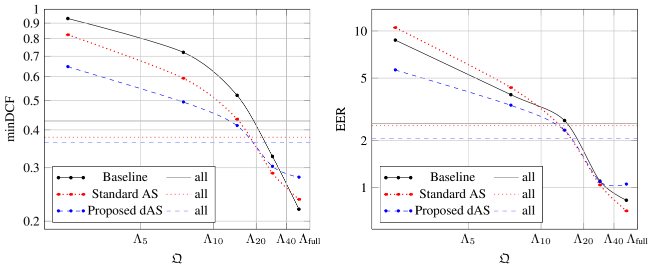
Figure 3: Log-normal distributed sample durations of development set data with respect to i-vector sufficiency classes

The vast majority of development samples are located in Λ_{full} (34.7%), Λ_{40} (31.1%), and Λ_{20} (23.1%), then: Λ_{10} (9.0%), and Λ_5 (2.1%). Intentional, all i-vectors have been centralized to the origin by mean-subtraction in the preparation of

[2] On each validation run one enrollment i-vector was randomly taken as a probe while the remaining i-vectors were used to create a template.

[3] The usage of information about other trials within the evaluation data is not allowed by the NIST challenge protocol [19].

[4] NIST set the i-vector challenge operating point similar to NIST SRE'10 at an effective prior $\tilde{\pi} = \frac{1}{101}$ [19, 24] with a Bayes threshold of $\eta \approx 4.6$.



(a) MinDCF

(b) EER

Figure 4: Biometric performance of i-vector sufficency classas: (a) minDCF, (b) EER.

the baseline system, but an unpaired Student t-test of independence showed that i-vector elements have significantly different mean-values compared between all development set i-vectors and with respect to each sufficiency class. Table 2 compares the amount of significantly independent i-vector elements according to their sample durations assuming equal variance[5].

Table 2: Student t-test of independent i-vector elements with respect to sufficiency classes

dev-set	all	Λ_{full}	Λ_{40}	Λ_{20}	Λ_{10}
Λ_5	84	141	91	66	44
Λ_{10}	142	230	140	70	
Λ_{20}	132	246	118		
Λ_{40}	35	180			
Λ_{full}	172				

Once mean-subtraction and whitening has been applied, Λ_{40} i-vectors exhibit the lowest significant offset to the space origin by having the second most impact on both i-vector processing due to their representative amount. Further, the most-sufficient i-vectors have the greatest gap compared to all development set i-vectors and to each sufficiency class with at least $140/600$ significant different mean positions. Hence, within the subspace of Λ_{40} i-vectors seem to be between the subspaces of high-insufficient and high-sufficient i-vectors. An opposite effect could be observed on short-duration samples, where the according i-vectors have larger mean-differences to i-vectors of more than 20 seconds than to i-vectors of comparable short duration samples (less than 20 seconds). This effect may be caused due to high variability of insufficient estimated i-vectors of short-duration samples, *i.e.* i-vectors of less than 20 second samples are distributed in subspaces that are more close to themselves than to more-sufficiently estimated i-vectors.

That is, offset vectors can be assumed for each sufficiency group, which effect the cosine score values due to angle changes between i-vectors[6]. These facts underline the need for compensating scoring statistics with respect to sample durations.

4.3. Performance Evaluation

Focusing on the baseline system the highest performance loss in terms of minDCF is observed for low-durational samples, see Table 3. As it can be seen, Λ_5 i-vectors yielded the most expensive detection costs with 0.932 which is very close to a random recognizers performance of $minDCF = 1$. I-vectors stemming from the class with the longest sample duration yielded the best observed minDCF, *i.e.* 0.219. However, on all other quality classes of insufficient i-vectors both AS-normalizations yield significant gains where the duration-invariant AS-norm outperforms both other systems on samples shorter than 20 seconds. On 20–40 second samples both normalizations could outperform the baseline approach, where AS-norm without duration-sensitive extension achieved the best minDCF for Λ_{40} i-vectors. Hence, AS-norm is necessary on insufficiently estimated i-vectors, and the proposed duration-based extension can yield up to 19.1% more relative-gain than the standard AS-norm.

Table 3: Duration group performances: avg. minDCF

System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.932	0.721	0.520	0.327	**0.219**
AS-norm	0.824	0.592	0.434	**0.288**	0.236
dAS-norm	**0.646**	**0.494**	**0.413**	0.303	0.279

In terms of biometric recognition performance both AS-norm approaches outperform the baseline as well, see Table. 4. Again, the proposed duration-invariant AS-norm yields significant gains on samples shorter than 20 seconds on which a performance break-down for the standard AS-norm can be observed. However, on higher-sufficient i-vectors the standard AS-norm outperforms both other systems, which could be caused due to the non-duration-invariance within the Λ_{full}

[5] Results of an unpaired Student t-test assuming un-equal variances yielded negligible differences in the results.

[6] Which actually is additive noise that should be well-compensable by, *e.g.* GPLDA scoring.

i-vectors. EER and minDCF performance comparisons among quality classes Ω are shown in Fig. 4.

Table 4: Duration group performances: avg. EER

System	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	8.74	3.92	2.68	1.10	0.83
AS-norm	10.51	4.35	**2.32**	**1.04**	**0.71**
dAS-norm	**5.63**	**3.35**	**2.32**	1.09	1.05

Across the entire set of classes the proposed duration-based AS-norm outperforms both other systems, see Table 5. In summary, the proposed duration-invariant AS-norm yields a 19.5% relative-gain in EER, a 32.6% relative-gain in FMR100, and a 15.0% relative-gain in minDCF compared to the baseline system on the cross-validation. Further, the duration-invariant AS-norm significantly outperforms the standard AS-norm which can also be seen in Fig. 5.

Table 5: System performances: avg. EER, FMR100, minDCF

System	EER	FMR100	minDCF	Challenge[7]
Baseline	2.56	5.15	0.428	0.386
AS-norm	2.49	4.48	0.378	0.331
dAS-norm	**2.06**	**3.47**	**0.364**	**0.312**

The results were approved by the preliminary evaluation of the 2013–2014 NIST i-vector challenge, where the application of the standard AS-norm resulted in a 14.2% relative-gain, and the duration-invariant extension resulted in a 19.2% relative-gain in minDCF.

Fig. 5 compares the best cross-validation systems according to minDCF within a Detection Error Trade-off diagram. The duration-invariant AS-norm improves the biometric performance of the baseline system at all operating points, while the standard AS-norm mainly yields gains in high-secure regions, *i.e.* operating points at low FMRs. In this regions both AS-normalizations exhibit equal recognition accuracy.

Hence, the proposed duration-invariant AS-norm extension is applicable to a larger range of scenarios compared to the standard AS-norm. While the duration-invariant AS-norm only obtains slightly lower error-rates on minDCF-operating points compared to the standard AS-norm, another advantage of the duration-invariant treatment is observed within entropy evaluations.

4.4. Entropy Evaluation

Table 6 compares the total C_{llr} of the three systems over all scores, and among each quality class. On Λ_5 i-vectors the baseline and the standard AS-norm perform similar to or worse than a random recognizer, and on samples having more than 5 seconds the standard AS-norm significantly outperforms the baseline system. On high-sufficient i-vector the lowest application-independent entropy was measured for the standard AS-norm with $C_{\text{llr}} = 0.05$, representing a very low cost of the LLR-scores. However, on sample durations lower than 40 seconds the duration-invariant AS-norm outperforms both other approaches by yielding a maximum LLR cost of $C_{\text{llr}} = 0.35$ on high-insufficient Λ_5 i-vectors. Overall the suggested AS-norm extension exhibits the lowest application-independent entropy by yielding relative-gains of 88.8% and 41.2%, respectively.

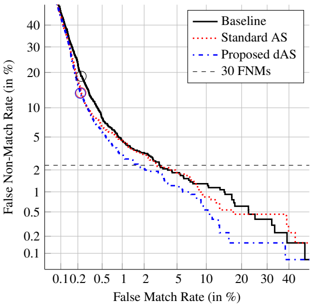


Figure 5: Systems detection error tradeoff: best systems from 10 cross-validations according to their minDCF.

Table 6: Average entropy comparison: all scores & duration-groups

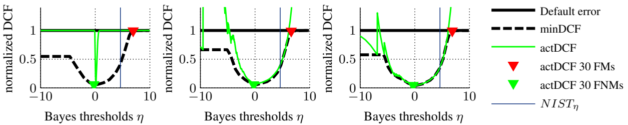
System	all	Λ_5	Λ_{10}	Λ_{20}	Λ_{40}	Λ_{full}
Baseline	0.89	0.95	0.93	0.92	0.89	0.86
AS-norm	0.17	1.18	0.41	0.18	0.08	**0.05**
dAS-norm	**0.10**	**0.35**	**0.20**	**0.11**	**0.07**	0.07

Fig. 6 illustrates the C_{llr} gains on normalized DCFs or likewise normalized Bayesian entropy plots, where the actual DCF (actDCF) represents application-dependent entropy, and the minDCF represents application-dependent entropy on a well-calibrated system — in these terms C_{llr} represents the area under actDCF, since we want to place emphasize on robustness, *i.e.* systems which do not require score-calibration. Due to the cosine scoring most scores of the baseline system lie within the range $[-1, +1]$, hence, the lowest DCF. The smallest difference between actual and minimum DCF was observed on $\eta \approx 0$, on any other operating point the baseline system is effected by huge mis-calibrations. Calibration-improvements were gained by the standard AS-norm, which delivers adequate calibration for a different application-points (actDCF curve being equal to minDCF curve). However, the suggested duration-invariant score-normalization yields well-calibrated scores on the vast majority of application-points, which have significant error-rates. That is, the proposed duration-invariant enables an enhanced statistical treatment of quality classes, which is approved by a very low overall entropy emission in terms of C_{llr}.

4.5. Discussion

Quality classes of i-vector sufficiency were motivated by assuming acoustical holes depending on the logarithmic sample duration. By observing i-vector mean offsets between the qual-

[7] The results were obtained by the 2013–2014 NIST i-vector online leaderboard which comprised 40% of the total evaluation data.



(a) Baseline (b) Standard AS-norm (c) Proposed dAS-norm

Figure 6: Entropy comparison of (a) the baseline system, (b) the standard AS-norm and (c) the proposed duration-invariant AS-norm.

ity classes $\mathfrak{Q}$, the need of duration-invariant recognition processes were empirically motivated in order to compensate for i-vector subspace mismatches. Hence, the AS-norm was extended with respect to duration-based quality classes as proposed in Sect. 3.3.

The experimental results showed that statistical effects of acoustical holes causing entropy are easy-compensable by analyzing their i-vector subspace variations according to same-shaped quality classes. Hence, additional processing-entropy was prevented for many operating points, and significant performance gains were yielded on short-duration samples as well. However, on high-sufficient i-vectors the standard AS-norm provides slightly better results, thus combined systems are considered promising with respect to recognition performance. Furthermore, more detailed separation of quality classes within Λ_{full} are expected to yield further gains within the proposed duration-invariant AS-norm.

Placing emphasize on the computational complexity the standard AS-norm requires all 36 572 development set i-vectors for either of the z-norm and t-norm sets in order to determine the top100 cohorts. In contrast, the duration-invariant extension utilizes at most 34.7% of the data amount for z-norm for Λ_{full} quality class normalizations, and 19.4% of the complete development set. Thus, proposed extension turns out to be highly suitable to units having less computational resources.

5. Conclusion

The proposed duration-invariant extension of AS-norm is proven to be highly suitable for applications in numerous use cases regarding industry as well as forensic, since it exhibits high performance by robust evidence strength. Hence, entropy in short-term duration classes could be reduced significantly, and a 19% relative-gain in biometric performance can be observed compared to the baseline system, proving the soundness of the presented approach. Further, the overall forensic evidence strength could be significantly increased, reducing the actual LLR cost to $C_{\text{llr}} = 0.10$.

Building upon our reproducible technique, future research might investigate template and probe normalizations as performed by Eq. 7 where quality-class-dependent mean offsets are vanished to achieve higher recognition performances. Hence, more sufficient comparators such as GPLDA or the two-covariance model [26], GPLDAs dot-product variation for fast scoring, can be applied on low-entropy i-vectors to concern more signal-based rather than processing-based entropy.

Further, duration-based quality classes can be investigated on invariant treatments towards their specific characteristics as i-vector subspaces on earlier processing stages such as duration-based i-vector extraction techniques.

6. Acknowledgement

This work has been partially funded by the European FP7 FIDELITY project (SEC-2011-284862) and the Center for Advanced Security Research Darmstadt (CASED).

7. References

[1] A. K. Jain, A. Ross, and S.Prabhakar, “An Introduction to Biometric Recognition,” in *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 2004.

[2] J. A. Batchelor, D. M. Lee, D. P. Banks, D. A. Crosby, K. W. Moore, S. H. Kuhn, T. Rodriguez, and A. B. Stephens, “Ivestigative report,” Florida Department of Law Enforcement, 2012, Laboratory report on US law case: State of Florida v. George Zimmerman.

[3] Dana Averbouch and Jade Kahn, “Fraud Targets the Contact Center: What Now?,” Speech Technology Magazine, November 2013, White paper of NICE systems.

[4] SESTEK, “The Rise of Voice Biometrics as a Key Security Solution,” Speech Technology Magazine, July 2013, White paper of SESTEK.

[5] T. Kinnunen and H. Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors,” in *EURASIP Speech Communication*, 2010.

[6] A. Fazel and S. Chakrabartty, “An Overview of Statistical Pattern Recognition Techniques for Speaker Verification,” in *IEEE Circuits and Systems Magazine*, 2011.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2011.

[8] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “Evaluation of i-vector Speaker Recognition Systems for Forensic Applications,” in *ISCA Interspeech*, 2011.

[9] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “i-vector Based Speaker Recognition on Short Utterances,” in *ISCA Interspeech*, 2011, pp. 2341–2344.

[10] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, “Duration Mismatch Compensation for i-Vector based Speaker Recognition Systems,” in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2013.

[11] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, “Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions,” in *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[12] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, “Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification,” in *ISCA Interspeech*, 2012.

[13] N. Fatima and T. F. Zheng, “Short Utterance Speaker Recognition — A research Agenda,” in *IEEE International Conference on Systems and Informatics (ICSAI)*, 2012.

[14] C. Zhang, X. Wu, T. F. Zheng, L. Wang, and C. Yin, “A K-Phoneme-Class based Multi-Model Method for Short Utterance Speaker Recognition,” in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[15] T. Stadelmann and B. Freisleben, “Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition,” in *International Conference on Pattern Recognition*, 2010.

[16] V. Hautamäki, Y.-C. Cheng, P. Rajan, and C.-H. Lee, “Minimax i-vector extractor for short duration speaker verification,” in *ISCA Interspeech*, 2013.

[17] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector Length Normalization in Speaker Recognition Systems,” in *ISCA Interspeech*, 2011.

[18] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, “Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System,” in *Iberoamerican Congress on Pattern Recognition (CIARP)*, 2013.

[19] C. Greenberg, “The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge,” 2013.

[20] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, “A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation,” in *ISCA Interspeech*, 2013.

[21] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *Odyssey*, 2010.

[22] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine Similarity Scoring without Score Normalization Techniques,” in *Odyssey*, 2010.

[23] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*, International Organization for Standardization and International Electrotechnical Committee, Mar. 2006.

[24] N. Brümmer, “The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF,” 2011.

[25] N. Brümmer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” in *ISCA Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[26] S. Cumani, N. Brümmer, L. Burget, and P. Laface, “Fast discriminative speaker verification in the i-vector space,” in *ICASSP*, 2011.

BIBLIOGRAPHY

[1] A. K. Jain, A. Ross, and S.Prabhakar, "An Introduction to Biometric Recognition," in *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 2004.

[2] J. A. Batchelor, D. M. Lee, D. P. Banks, D. A. Crosby, K. W. Moore, S. H. Kuhn, T. Rodriguez, and A. B. Stephens, "Ivestigative report," Florida Department of Law Enforcement, 2012, laboratory report on US law case: State of Florida v. George Zimmerman.

[3] D. Averbouch and J. Kahn, "Fraud Targets the Contact Center: What Now?" Speech Technology Magazine, November 2013, white paper of NICE systems.

[4] SESTEK, "The Rise of Voice Biometrics as a Key Security Solution," Speech Technology Magazine, July 2013, white paper of SESTEK.

[5] S. Hegenbart, "Sprecherverifikation — Implementierung eines Verifikationssystems unter Verwendung der Joint Factor Analysis," Master's thesis, Hochschule Mannheim, 2013.

[6] S. Billeb, "Template Protection für biometrische Sprecherverifikation nach ISO/IEC 24745," Hochschule Darmstadt, University of Applied Science, atip GmbH, 2014, B.Sc. thesis.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[8] L. R. Rabiner, "A Tutorial on Hidden-Markov-Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, 1989.

[9] E. G. Schukat-Talamazzini, *Automatische Spracherkennung — Statistische Verfahren der Musteranalyse*. Vieweg Verlag, 1995.

[10] A. Martin, "The 2009 NIST Language Recognition Evaluation Plan (LRE09)," 2009, last viewed on 29.11.2013. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf

[11] L. J. Rodríguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE),"

2012, last viewed on 29.11.2013. [Online]. Available: http://www.ehu.es/~ljrf/tmp/Albayzin_LRE12_EvalPlan_v1.2.pdf

[12] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," in *EURASIP & ISCA Speech Communication*, 2010.

[13] A. Fazel and S. Chakrabartty, "An Overview of Statistical Pattern Recognition Techniques for Speaker Verification," in *IEEE Circuits and Systems Magazine*, 2011.

[14] H. Baier, F. Freiling, and B. Roos, "Einführung in die Computer Forensik," 2012, Hochschule Darmstadt, University of Applied Science.

[15] J. Mortera and A. P. Dawid, "Probability and Evidence," Department of Statistical Science, University College London, Tech. Rep., 2006, Research Report No. 264, last viewed on 15.11.2013. [Online]. Available: http://www.ucl.ac.uk/statistics/research/reports

[16] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition," in *ISCA Eurospeech*, 2003.

[17] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector Speaker Recognition Systems for Forensic Applications," in *ISCA Interspeech*, 2011.

[18] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, University of Stellenbosch, 2010.

[19] ——, "The Role of Proper Scoring Rules in Training and Evaluating Probabilistic Speaker and Language Recognizers," ISCA Speaker Odyssey, 2012, slides.

[20] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLbook.com, 2012.

[21] ISO/IEC, "Information technology — Vocabulary — part 37: Biometrics," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 23827:2012(E), 2012, Harmonised biometric vocabulary.

[22] F. Kelly, N. Brümmer, and N. Harte, "Eigenageing Compensation for Speaker Verification," in *ISCA Interspeech*, 2013.

[23] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," in *IEEE Transacions on Audio, Speech, and Language Processing*, 2007.

[24] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *ISCA Speaker Odyssey*, 2010.

[25] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[26] ISO/IEC, "Information technology – Biometric performance testing and reporting – Part 1: Principles and framework," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19795-1:2006(E), 2011.

[27] ——, "Text of Standing Document 11, Part 1 Harmonization Document," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC, 2010.

[28] ——, "Information technology — Biometric application programming interface — Part 1: BioAPI specification," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19784-1:2006(WD), 2006.

[29] ——, "Information technology — Biometric data interchange formats — Part 1: Framework," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19794-1:2011(E), 2011.

[30] ——, "Information technology — Security techniques — Biometric information protection," JTC 1/SC 27, Geneva, Switzerland, ISO/IEC 24745:2011(E), 2011.

[31] ——, "Multi-Modal and Other Multi-Biometric Fusion," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 24722:2007(TR), 2007.

[32] ——, "Information technology – Biometric performance testing and reporting – Part 2: Testing methodologies for technology and scenario evaluation," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19795-2:2007(E), 2011.

[33] T. Fawcett, "An introduction to ROC analysis," in *IAPR Pattern Recognition Letters*, 2006.

[34] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *ISCA Eurospeech*, 1997.

[35] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective," in *EURASIP & ISCA Speech Communication*, 2000.

[36] B. D. Jovanovic and P. S. Levy, "A Look at the Rule of Three," in *ASA The American Statistician*, 1997.

[37] P. D. Grünwald and A. P. Dawid, "Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory," in *ims The Annals of Statistics*, 2004, last viewed on 14.11.2013. [Online]. Available: http://www.ucl.ac.uk/statistics/research/reports

[38] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," in *Journal of the American Statistical Association (ASA)*, 2007.

[39] G. W. Brier, "Verification of forecasts expressed in terms of probability," in *AMS Monthly Weather Review*, 1950.

[40] N. Brümmer, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," 2011, last viewed on 24.01.2014. [Online]. Available: http://arxiv.org/pdf/1304.2865v1.pdf

[41] ——, "Application-Independent Evaluation of Speaker Detection," in *ISCA Speaker Odyssey*, 2004.

[42] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," in *ISCA Speaker Odyssey*, 2006.

[43] N. Brümmer, "Calibration of Likelihood-Ratios in Automatic Speaker Recognition," BBfor2 Workshop, 2011, slides.

[44] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," in *Conversational Speech, Digital Signal Processing*, 2000.

[45] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, "Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System," in *Iberoamerican Congress on Pattern Recognition (CIARP)*, 2013.

[46] C. Greenberg, "The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge," NIST, Tech. Rep., 2013, last viewed on 11.10.2013. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

[47] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[48] D. A. Reynolds, "Gaussian Mixture Models," MIT Lincoln Laboratory, Tech. Rep., 2007.

[49] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds." in *msp Pacific Journal of Mathematics*, 1968.

[50] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," in *IEEE Transaction on Audio, Speech, and Language Processing*, 2008.

[51] P. Kenny, N. Dehak, V. Gupta, P. Ouellet, and P. Dumouchel, "A new training regimen for factor analysis of speaker variability," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2008.

[52] O. Glembek, "Joint Factor Analysis Matlab Demo," 2009, last viewed on 10.10.2013. [Online]. Available: http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo

[53] Y. Singer and M. K. Warmuth, "Batch and On-line Parameter Estimation of Gaussian Mixtures Based on the Joint Entropy," in *Advances in Neural Information Processing Systems*, 1999.

[54] R. A. Stine, "Model Selection using Information Theory and the MDL Principle," 2003, The Wharton School of the University of Pennsylvania, Department of Statistics.

[55] The MathWorks, Inc., "Gmdistribution class," 2012, Matlab Documentation.

[56] ——, "Rankfeatures function," 2012, Matlab Documentation.

[57] C. Longworth, "Kernel Methods for Text-Independent Speaker Verification," Ph.D. dissertation, Cambridge University Engineering Department and Christ's College, 2010.

[58] The MathWorks, Inc., "Classification Using Nearest Neighbors," 2012, Matlab Documentation.

[59] P. J. Olver, "Numerical Analysis Lecture Notes," 2008, last viewed on 07.01.2014. [Online]. Available: http://www.math.umn.edu/~olver/num_/lnv.pdf

[60] R. W. Picard, "Topic: Decorrelating and then Whitening data," 2008, notes, last viewed on 07.01.2014. [Online]. Available: http://courses.media.mit.edu/2010fall/mas622j/whiten.pdf

[61] M. Welling, "Fisher Linear Discriminant Analysis," 2009, last viewed on 07.01.14. [Online]. Available: http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf

[62] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality Reduction: A Comparative Review," Tilburg University, Tech. Rep., 2009, tiCC-TR 2009-005, last viewed on 10.10.2013. [Online]. Available: http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html

[63] S. Cumani, N. Brummer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i-vector space," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2011.

[64] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine Similarity Scoring without Score Normalization Techniques," in *ISCA Speaker Odyssey*, 2010.

[65] M. Honda, "Human Speech Production Mechanisms," in *NTT Technical Review*, 2003.

[66] B. Pompino-Marschall, *Einführung in die Phonetik*, ser. de Gruyter Studienbuch. Walter de Gruyter, 2003.

[67] Rosistem Barcode, "Biometric Education — Voice Recognition," 2003, last viewed on 27.02.2014. [Online]. Available: http://www.barcode.ro/tutorials/biometrics/voice.html

[68] H. Robjohns, "A brief history of microphones," in *Microphone Data Book*, 2010.

[69] C. Greenberg, "The NIST Year 2012 Speaker Recognition Evaluation Plan," NIST, Tech. Rep., 2012, last viewed on 11.10.2013. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

[70] B. Ziółko and M. Ziółko, "Time Durations of Phonemes in the Polish Language," in *Language & Technology Conference: Human Language Technology*, 2009.

[71] C. Bagwell, "Sox – sound exchange, the swiss army knife of audio manipulation," 2013, manual page, last viewed on 21.02.2014. [Online]. Available: http://sox.sourceforge.net/sox.html

[72] P. Rose, *Forensic Speaker Identification*. Taylor & Francis, Forensic Science Series, 2002.

[73] J. Rajnoha and P. Pollák, "ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness," in *Radioengineering*, vol. 20, 2011, pp. 74–84.

[74] N. Brümmer, A. Swart, L. Burget, S. Cumani, O. Glembek, M. Karafiát, P. Matejka, O. Plchot, M. Soufifar, J. Silovský, P. Kenny, J. Alam, P. Dumouchel, P. Ouellet, M. Senoussaoui, and T. Stafylakis, "ABC System description for NIST SRE 2012," Agnitio, BUT (Technical University of Liberec), CRIM, Tech. Rep., 2012.

[75] D. Colibro, C. Vair, K. Farrell, N. Krause, K. Gennady, S. Cumani, and P. Laface, "Nuance — Politecnico di Torino (NPT) System Description for NIST 2012 Speaker Recognition Evaluation," Nuance Communications & Politecnico di Torino, Tech. Rep., 2012.

[76] M. Diez, M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "University of the Basque Country Systems for the NIST 2012 Speaker Recognition Evaluation," University of Basque Country, Tech. Rep., 2012, NIST SRE'12 Workshop.

[77] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *ISCA Speaker Odyssey — The Speaker Recognition Workshop*, 2001.

[78] J. Gonzalez-Dominguez, J. Franco-Pedroso, D. Ramos, D. T. Toledano, J. Gonzalez-Rodriguez, A. Kanagasundaram, D. Dean, and S. Sridharan, "ATVS-QUT NIST SRE 2012 System Description," in *NIST SRE'12 Workshop*, 2012.

[79] T. Hasan and J. H. L. Hansen, "A study on Universal Background Model training in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1890–1899, 2011.

[80] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling With Sparse Training Data," in *IEEE Transactions on Speech and Audio Processing*, 2005.

[81] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Centre de recherche informatique de Montréal (CRIM), Tech. Rep., 2005.

[82] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *ISCA Interspeech*, 2011.

[83] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation," in *ISCA Interspeech*, 2013.

[84] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE International Conference on Computer Vision ICCV*, 2007, last viewed on 25.10.2013. [Online]. Available: http://www.cs.ucl.ac.uk/staff/s.prince/Papers/ICCV2007PLDAFinal2.pdf

[85] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA Modeling in i-Vector and Supervector Space for Speaker Verification," in *ISCA Interspeech*, 2012.

[86] P. Matějka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. H. Černocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2011.

[87] J. Villalba and N. Brümmer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance," in *ISCA Interspeech*, 2011.

[88] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition Systems," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2013.

[89] L. Ferrer, Y. Lei, M. McLaren, N. Scheffer, M. Graciarena, and V. Mitra, "SRI 2012 NIST Speaker Recognition Evaluation System Description," Stanford Research Institute (SRI), Tech. Rep., 2012.

[90] L. Ferrer, A. Lawson, Y. Lei, M. McLaren, and N. Scheffer, "SRI Submission for NIST SRE 2012," NIST SRE'12 Workshop, 2012.

[91] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances," in *ISCA Interspeech*, 2011, pp. 2341–2344.

[92] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," in *ISCA Interspeech*, 2012.

[93] R. Vogt, S. Sridharan, and M. Mason, "Making Confident Speaker Verification Decisions with Minimal Speech," in *ISCA Interspeech*, 2008.

[94] R. Vogt and S. Sridharan, "Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions," in *ICB Proceedings of the Third International Conference on Advances in Biometrics*, 2009.

[95] C. Greenberg, "The NIST Year 2010 Speaker Recognition Evaluation Plan," NIST, Tech. Rep., 2010, last viewed on 11.10.2013. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

[96] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, "i-Vectors in the Context of Phonetically-constrained short utterances for Speaker Verification," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2012.

[97] N. Fatima and T. F. Zheng, "Short Utterance Speaker Recognition — A research Agenda," in *IEEE International Conference on Systems and Informatics (ICSAI)*, 2012.

[98] C. Zhang, X. Wu, T. F. Zheng, L. Wang, and C. Yin, "A K-Phoneme-Class based Multi-Model Method for Short Utterance Speaker Recognition," in *APSIPA Annual Summit and Conference (ASC)*, 2012.

[99] R. Vogt, B. Baker, and S. Sridharan, "Factor Analysis Subspace Estimation for Speaker Verification with Short Utterances," in *ISCA Interspeech*, 2008.

[100] H. Lei, "Joint Factor Analysis (JFA) and i-vector Tutorial," 2011, last viewed on 12.03.2014. [Online]. Available: http://www1.icsi.berkeley.edu/Speech/presentations/AFRL_ICSI_visit2_JFA_tutorial_icsitalk.pdf

[101] Linguistic Data Consortium (University of Pennsylvania), "1996-2008 NIST Speaker Recognition Evaluation Data Collection," Linguistic Data Consortium (University of Pennsylvania), 2009, Multilingual (English, Spanish, Arabic, Chinese, Russian, Tagalog, Korean, unknown), catalog number: LDC2009E100.

[102] ——, "SRE12 training audio data from SRE10," Linguistic Data Consortium (University of Pennsylvania), 2012, Multilingual (English, Spanish, Arabic, Chinese, Russian, Tagalog, Korean, unknown), catalog number: LDC2012E09.

[103] ——. (2012) SRE12 Evaluation audio data. Multilingual (English, Spanish, Arabic, Chinese, Russian, Tagalog, Korean, unknown), catalog number: LDC2012Evaluation.

[104] J. Borgstrom, W. Campbell, N. Dehak, R. Dehal, D. Garcia-Romero, K. Greenfield, A. McCree, D. Reynolds, F. Richardson, E. Singer, D. Sturim, and P. Torres-Carrasquillo, "MITLL 2012 Speaker Recognition Evaluation System Description," Massachusetts Institute of Technology Lincoln Laboratory and Johns Hopkins University Human Language Technology Center of Excellence, Tech. Rep., 2012.

[105] S. Strassel, K. Jones, D. Graff, K. Walker, J. Wright, and C. Cieri, "Linguistic Resources for the NIST SRE12 Evaluation," 2012, slides from the 2012 NIST SRE workshop.

[106] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," in *EURASIP Journal on Applied Signal Processing*, 2004.

[107] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkenafet, L. M. Fernández, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Günther, J. Žganec Gros, R. Z. Candil, F. S. oes, M. Bengherabi, A. Álvarez Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz, D. van Leeuwen, J. González-Domínguez, M. U. Neto, E. Boutellaa, P. G. Vilda, A. Varona, D. Petrovska-Delacrétaz, P. Matějka, J. González-Rodríguez, T. Pereira, F. Harizi, L. J. Rodriguez-Fuentes, L. E. Shafey, M. Angeloni, G. Bordel, G. Chollet, and S. Marcel, "The 2013 Speaker Recognition Evaluation in Mobile Environment," Idiap Research Institute, Alpineon Ltd., Universidad Autónoma de Madrid, CPqD, Centre de Développement des Technologies Avancées (DZ), Universidad Politécnica de Madrid, University of the Basque Country, L2F/INESC-ID (PT), Institut Mines-Télécom, Phonexia s.r.o., Radboud University Nijmegen, Brno University of Technology, Tech. Rep., 2013, last viewed 13.10.2013. [Online]. Available: http://publications.idiap.ch/downloads/papers/2013/Khoury_ICB2013_2013.pdf

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URWPalladio L* and *FPL*) are used. The "typewriter" text is typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera".

The diagrams in this thesis were created using *TIKZ*, *PGFPlots* by Christian Feuersänger, *Microsoft Office Visio 2007*, *Matlab R2012a*, the *BOSARIS* toolkit by Niko Brümmer (Agnitio Labs), *matlab2tikz* by Nico Schlömer, *matlabfig2pgf* by Paul Wagenaars, and *rastamat* by Daniel P. W. Ellis.

Final Version as of April 1, 2014 (`classicthesis` camera-ready).