Impostor-centric T-norm in speaker verification.

Pearson, Neil

Impostor-centric T-Norm in speaker verification

by

Neil Pearson

Supervisor: Dr J.S.D. Mason

A thesis submitted to the
University of Wales
in fulfilment for the degree of
MASTER OF PHILOSOPHY

School of Engineering

SWANSEA UNIVERSITY

2008

Swansea University
Prifysgol Abertawe

Abstract

Normalisation is key to performance in speaker recognition. Such approaches conducted in the score domain attempt to alleviate inter-speaker perturbations. Published results using the conventional test-normalisation (T-Norm) approach show large performance variations when different cohort compositions are used to derive the score normalisation statistics. Two issues arise with the use of background normalisation speakers: a selection regime and its quantity. This thesis illustrates the variability of a robust speaker verification system when selecting such cohorts. An empirical investigation is conducted on the popular T-Norm approach with a variety of selection procedures to compose the impostor cohort. Prior knowledge is extra information that can be utilised to focus selection. The quantity of impostors is an important attribute. For example, when computational or storage resources are a premium, a cohort containing as little as 15 models is able to provide admirable verification performance when a certain selection criterion is utilised. Results also show different performances with the individual mean and standard deviation components of the distribution scaling approach utilised by the T-Norm. A higher sensitivity is shown by the mean only component during shorter utterance evaluations, however, the degrading effect of the mean component is reduced with enhanced training duration. Such results are introduced with the conventional T-Norm approach, here deemed as a *trial-independent* approach, as the same cohort is applied to all trials in a given evaluation.

The *trial-specific* scenario is a further variant of T-Norm. Here, each target speaker is provided with a personal selection of impostor models to represent a normalisation cohort. One such procedure is the adaptive T-Norm (AT-Norm), known to giving general improvement over the conventional T-Norm. However, a large pool of potential impostors is required to supply a range of target comparative impostor models for the potential variety of speaker enrolment.

The final part of the work extends on the investigations undertaken on T-Norm. Improvement can be shown when the influence of a few poorly derived target models containing little speaker discrimination are reduced. Denoted as the *speaker security measure* (SSM), this process highlights models of a poor enrolment quality which can be addressed accordingly. Here, it is shown that a simple weighting procedure reduces the influence of poorly trained models whilst enhancing models deemed of higher discrimination. Improvement can be shown when the influence of a few poorly derived target models containing little speaker discrimination are reduced.

Very late in this work, NIST released, at the end of March 2008, a new revision of the 2006 evaluation trial keys. The thesis contains a short epilogue presenting results on the 10sec-10sec and 1conv-1conv with the revised and original keys. This highlights the influence of database correctness when conducting the sort of evaluations central to the research reported in this thesis.

Declaration/Statements

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ... . (N. Pearson, candidate)

Date ..... 29/5/08 .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (N. Pearson, candidate)

Date ..... 29/5/08 .....

Statement 2

I hearby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (N. Pearson, candidate)

Date ..... 29/5/08 .....

Acknowledgements

The period of my research study and hence the completion of this thesis would not have been possible without the steady support of many special people, hopefully, most will be listed below.

First, I would like to thank my supervisor Dr J.S.D. Mason for guidance, enthusiasm and the numerous discussions in the field of data-driven classification. I would also like to thank Dr Nick Evans for counselling in the world of Unix during the early stages of my research.
I would like to thank all associates of the speech and image research group, in no particular order, Keith, Jasmine, Alicia, Benoit, Ruben and Richard for their friendship, assistance and contributions throughout my research period.

I gratefully acknowledge sponsorship by Silverwing UK Ltd with additional support from the employees and its managing director Simon Packer for giving me the opportunity, financial support, resources and work flexibility during my candidature.

An especial acknowledgment goes to my wife Siân, who's love, support and patience have been extraordinary during my periods of study.

My warmest thanks go to my parents Aneira and Dennis for their a range of advice, unlimited support throughout my education and their continual encouragement and guidance.

Finally, I would also like to thank all other supporting people that have inspired and encouraged me is some way shape or form, in particular Neal, Jonathan and Nadim for their friendship in both undergraduate and postgraduate studies.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

Abbreviation	Description	Chapter
SV	Speaker Verification	1.1
GMM	Gaussian Mixture Model	2.3
KL	Kullback-Leibeler divergence measure	4.1
SVM	Support Vector Machines	2.3
NAP	Nuisance Attribute Projection	2.3
DET	Detection Error Trade-off	2.8
EER	Equal Error Rate	3.1
PDF	Probability Density Function	2.2
MAP	Maximum A Posteriori	2.2
LLR	Log-likelihood Ratio	2.1
DCF	Detection Cost Function	3.1
NIST	National Institute of Standards and Technology	1.2
CMS	Cepstral Mean Subtraction	2.3
RASTA	RelAtive SpecTrA	2.3
SSM	Speaker Security Measure	5.2

Terminology

Term	Description
Z-Norm	Zero normalisation
C-Norm	Channel normalisation
H-Norm	Handset normalisation
T-Norm	Test normalisation
AT-Norm	Adaptive Test normalisation
D-Norm	Distance normalisation
True	The outcome of a verification has determined that two utterances come from the same person
False	The outcome of a verification has determined that two utterances come from different people
Trial	The individual evaluation of two signals
Test utterance	The utterance denoted as the unknown source for trial (B)
Target utterance	(or enrolment) The utterance, known or identified through an enrolment (A) procedure
Model	A representation of the speaker from a set of speech vectors by a collection of means, variances and weights
World model	A model created from out of class speech data
Universal Background Model	A model created from out of class speech data
Impostor Model	A speaker-specific model of trained on speech data that is deemed not from the target speaker during a trial
T-Norm pool	A collection of speaker models, in this thesis, primarily impostors
Selection	A procedure to select a cohort of models for T-Norm
Impostor-centric	A T-Norm pool that is constructed from a collection of impostors
Target-centric	A T-Norm pool composed of models representative of utterances from the target speaker
Evaluation set	A collection of trials where no speaker verification system development has been conducted
Development set	A collection of trials used to develop and attempt enhancement of a speaker verification system
Random T-Norm	A selection of speaker models used for all trials in an evaluation.
Matched T-Norm	When the speaker set is of similar representation as the target model
Miss-Matched T-Norm	A general group of models where the data set is not of the same representation as the target model
Trial-independent T-Norm	The same cohort of models that has been appointed to all trials in an evaluation, usually matched to the target criteria through some selective approach
Trial-specific T-Norm	A cohort of models that has similar attributes to a single trial, usually a single target, selected by a given criteria

Chapter 1

Introduction

1.1 Background

Speech is considered as a non-intrusive biometric that can be used to distinguish people. It is a biometric that can express an intent or identity claim whilst providing characteristics of a person through their speech. A practicable application scenario is a security enhancement for telephone banking. In some application scenarios, speech may be the only plausible biometric. Commercially, this is beneficial in the ubiquitous telephony and mobile phone as microphones are readily available, cheap and there is no need for special signal transducers [1]. A further overview of applications is discussed in [2].

Like all biometrics, a key cause of concern that can lead to errors is signal degradation. Examples of degradation nuance are environmental acoustics (babble, car, reverberation, etc.), transmission channel variation, physical handset properties, speaker health and phonetic content can all be embedded in the speech [1]. To quote Doddington [3] *"speech is a performing art and each performance is unique"*. Typical degradation introduces variations across sessions, referred to as inter-session variability and is considered a nuisance. Normalisation approaches attempt to tread the signal in all biometrics. Significant research has been undertaken to reduce these perturbation factors. This thesis is concerned with reducing these effects.

The assistance of additional sources of speaker knowledge can improve verification performance. Recognising people over the telephone by their speech is a trait that the majority of us use daily. Though the use of a 'Caller ID' display is a feature available on most phones today, it gives us an assumption of an identity before answering the call, a sense of *prior knowledge*[1]. Prior knowledge is discussed in [4] for use in different stages of the classifier in an attempt to reduce microphone miss-match between train and test utterances. The use of prior knowledge in text-independent speaker verification (SV) is commonly used in many systems. It allows for context dependent nor-

[1] Also regarded as *auxiliary* information from the NIST evaluation plans

malisation, for example, speaker gender.

Speaker recognition can be separated into two categories, identification and verification. SV attempts to confirm an unknown voice against a given claim for a particular system enrolled speaker; a one-to-one classification. Speaker identification is an n class problem with no claim provided. A ranking methodology can derive an identity of the unknown speaker by, for example, the highest score of the enrolled speaker set. The former context is the focus of this work.

1.2 Speaker Verification

The fundamental objective of a SV trial is to conclude the validity of two independent speech signals (labeled A & B) originating from the same person, asking the question *Does utterance B come from the same person as utterance A*? One utterance is often referred to as the training utterance, used for enrolment to produce a target model or true speaker [3]. The second utterance is conventionally captured subsequently in time from an unknown person and is usually defined as the test utterance[2]. For the benefit of the reader, the overuse of the word 'test' can become tied to the actual assessment of two utterances, causing confusion over the word *test*. Here an assessment of two utterances will be known as a *trial*.

Utterance A is conventionally provided with a label, albeit an assumed identity[3]. Utterance B denotes the unknown test utterance with a claim of being speaker A. Also note that the following experiments are conducted in accordance to the National Institute of Standards and Technology (NIST) evaluation protocol by performing all trials independently.

In real-world scenarios it is likely that both provided utterances have been recorded over different sessions under different conditions coupled with sources of degradation which are difficult to control. Different sources of signal degradation cause miss-matches between training and test data. Thomson [5] explains the ideal approach for classification is to use a matched ideology, showing that an impractical yet ideally matched condition arises when both training and test utterances originate from the same recording session as the condition of the speaker is unlikely to change in a short-time span. Variations in the speech are also established through the natural problem of speakers being unable to repeat an utterance precisely [6]. However, this is unrealistic and natural degradation, from such examples previously discussed, between the speech recordings is reflected throughout the classifier. There is major difficulty in reducing the effects of degraded signals; this concern leads to verification errors. Under these degrative conditions, we want to reduce these

[2]The test utterance can also be referred to as the *competing speaker*

[3]An identity is usually assumed for one utterance (here being A) but is not necessary as classification can be done in a bilateral manner, a reversal of roles. A model of A can be generated and evaluated against utterance B. The resulting scores are $P(A|\lambda_B)$ and $P(B|\lambda_A)$ which can be fused to usually enhance performance.

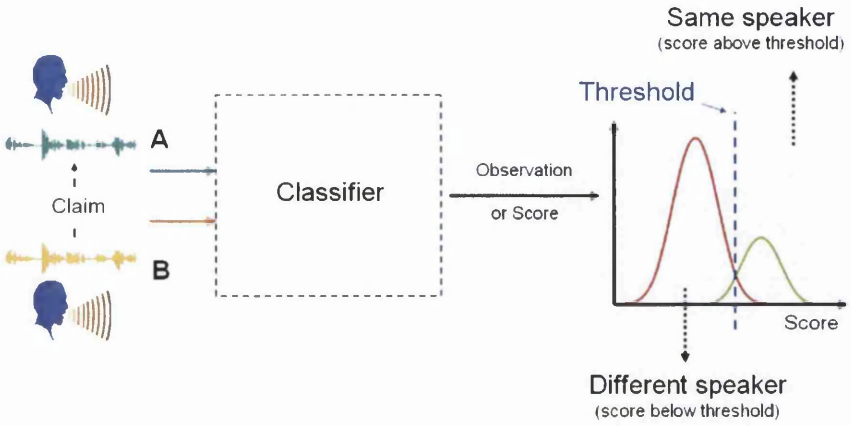


Figure 1.1: High-level speaker classifier view, generating a score. A decision is determined by the threshold, a score above dictates the same speaker and below states an impostor. The true (green) and false (red) allow the setting of the threshold from a prior evaluation.

inter-session variability. Another potential inter-session variability is the implication of the spoken text content may be different. This thesis focuses on ameliorating these conditions. In the previous paragraph, two components constitute inter-session variability, the speech content, e.g. text context an pronunciation of phonetic content, whilst the other variation is introduced by, for example, the environment. This thesis focuses on reducing the influence of the second class. Such processing can be done throughout the verification process, from feature extraction through to the decision process where each stage invariably involves some form of normalisation. In this thesis, the focus is on normalisation of the score or observation, shown in Figure 1.1. Outcomes are classified into true and false distributions of an evaluation set of many trials, depicted in Figure 1.1 as green and red distributions respectively. Setting a decision threshold that dictates a true or false classification is difficult for any scenario. Score normalisation assists this by reducing the variability of both resulting true and false distributions.

Normalising against this variability is referred to by Furui [6] as one of the most difficult problems in SV. Different forms or sources of normalisation are directed towards degradation. For example, speech may have passed through a communications system and incurred channel degradation; the handset normalisation (H-Norm) [7] approach, applied at the score level can reduce such effects. This implies prior knowledge about the channel which can be deduced by sub-classifiers or apriori information. Such knowledge can be gender, age, health [8], quantity and quality [9] of speech. All of which can contribute towards inter-session variability and therefore lead to degraded performance in the absence of any normalisation or compensation. Many of these parameters can

be known or determined as part of the trial exercise and the use of this prior knowledge within the classifier is central to the work described in this thesis. Specifically, information about the duration of speech is used in the score normalisation approach known as T-Norm [10]. The use of test-normalisation (T-Norm) has become commonplace within SV systems since it was first proposed in 2000 by Auckenthaler et al. [10]. Due to the availability of unspecific speaker utterances, this normalisation is usually achieved by applying additional sources of speaker related knowledge through the medium of speaker models, considered as impostors to the target trials. This is deemed a method of impostor-centric normalisation. Performance gains have been shown with T-Norm [10, 11] when applying a evaluation-dependent cohort to generate normalisation statistics. In brief, this is an extension to the matched methodology introduced by few participants in NIST 2004 and a more recent extension with an adaptive form of T-Norm (AT-Norm) [12]. For a matched scenario, the impostor models used to derive the normalisation statistics for T-Norm are *matched* by their enrolled quantity of speech to the target models over an evaluation. Adaptive T-Norm defines a personal cohort for each target in an evaluation.

For the work in this thesis, benefits of such matching are described for T-Norm and in particular the contributions of the mean and standard deviation are examined. Adaptive and conventional T-Norms are directly compared and in accordance to published results by Sturim and Reynolds [12] is found to give superior performances. This provides further enhancement to speaker-verification with well trained targets, though no publication has discussed the ramifications of AT-Norm on short-duration evaluations. The use of such approaches will be investigated and compared for both long and short-duration evaluations. In previous published work by Auckenthaler [10] and Sturim and Reynolds [12], the optimum number of impostor models is reported to be in the order of 50. Here we show that 15 for 10sec and 25 for 1conv[4] NIST conditions can be applied. These findings are believed to be new. The experimental work throughout this thesis are conducted in accordance with the NIST protocols, this and system configuration can be found in the appendix A for the convenience of the reader. In all cases, prior knowledge for the background model data for the background models and impostors come from NIST 2004, development on NIST 2005 and evaluations on NIST 2006. A final contribution in this thesis comes from the impostor score observations derived from T-Norm statistics which are hypothesised to indicate the robustness of the model against impostor attacks, a form of model quality. This is termed as the *speaker security measure* (SSM). This procedure can be considered as a pre-filter, using a strategy to either alleviate errors or reduce the influence of such speaker models which have been derived from non-speaker discriminative utterances.

[4]Utterance with approximately 2.5 to 3 minutes of speech. This is also interchangeable to *1side* or in the recent NIST 2008 evaluation plan, denoted as *short2*

1.3 Thesis Overview

The remainder of this thesis describes the general ideology of the classifier and published approaches to reduce signal degradation through normalisation. Work is then applied and presented at the score level to enhance the robustness of decision thresholds for both short and longer duration tasks. Experimental trials are conducted on the annual NIST speaker recognition evaluations. Experimental framework configuration and protocols are discussed in the appendix A.1 and A.2 respectively. The chapter outline of the thesis is as follows:

- Chapter 2 gives an overview of the three stages of a classification system with emphasis on published normalisation methods that can be applied to each module.
- The concept of cohort selection strategies in a trail-independent scenario for score normalisation is discussed, specifically for the T-Norm method is presented in Chapter 3. Experimental studies and results are discussed.
- Chapter 4 investigates the trial-dependent cohorts with examinations of two approaches to select impostors for T-Norm cohorts. Empirical observations are discussed when the AT-Norm is applied and contrasted to the conventional T-Norm approach.
- Model quality measures derived from the observations of the T-Norm scores are investigated in Chapter 5. The resultant confidence measure has been coined as the *speaker security measure* (SSM).
- Chapter 6 concludes the thesis and examines the potential of further investigations.

Chapter 2

Classification with Normalisation

In this chapter we follow from the general discussion of speaker verification (SV) by describing the classification tools used during subsequent experiment analysis. Primarily, this is concerned with the reduction of utterance perturbations through normalisation in the score domain that help reduce the effects of signal degradation. Here, the concept of using extra speaker, utterance or trial related information towards a SV trial is introduced under the umbrella of *prior knowledge*.

2.1 Speaker Verification

Speaker Verification (SV) has two possible outcomes of either true (the unknown speaker is judged to be the target) or false (the unknown speaker is identified as an impostor) generated by an observation classified by a pre-designed threshold θ; normally established empirically using a collection of development trials.
Observations concerning the claim of an unknown test speaker are derived from two hypothesis:

1. H_0 test utterance ***is*** from the hypothesis speaker

2. H_1 test utterance ***is not*** from the hypothesised speaker (the alternative hypothesis)

The context of H_1 is less well defined since it can potentially represent the entire space of possible alternatives. Again, utterance A and B are used to identify the conventional training and test utterances respectively.

Based on the binary hypothesis, two score are provided for a trial, $P(B|\lambda_A)$ is an observation of H_0 and $P(B|\overline{\lambda_A})$ denotes the score of the utterance under test (B) for H_1. λ denotes the target (speaker) model and $\overline{\lambda}$ denotes the alternative model. An observational likelihood score can then

be derived from these two hypothesis for an enrolled speaker λ_A and unknown utterance B .

$$H_0 \text{ is true when } \frac{P(B|\lambda_A)}{P(B|\overline{\lambda_A})} > \theta \tag{2.1}$$

$$H_1 \text{ is true when } \frac{P(B|\lambda_A)}{P(B|\overline{\lambda_A})} < \theta \tag{2.2}$$

Furui [6] refers to this as the similarity domain normalisation. The alternative model hypothesis increases discrimination by reducing additive effects (e.g. duration) of different test utterances [13] over many trials. Often the logarithm of the likelihood is used giving the log-likelihood ratio (LLR) , derived as equation 2.3, to reduce the deviations of high scoring trials. These are score observations used for subsequent experiments and verification decisions are derived against a prior threshold θ.

$$\Lambda(B)_T = \log(P(B|\lambda_A)) - \log(P(B|\overline{\lambda_{UBM}})) \tag{2.3}$$

For a text-independent classifier, the influence of the test utterance length has been reduced through normalisation, using the difference from both the target model and UBM.

A classification system decision threshold is set by utilising a development set of trials with assumed outcomes, denoted as *ground-truth.* Tailoring the threshold θ to a specific application is conducted by observing numerous trials from a development set inclusive of the assumed outcome of each trial conducted known ground-truths. To set a threshold, scores from a development set can be separated into two distributions using some ground-truth, the true and false distributions (green and red respectively in Figure 1.1). Conventionally, a score above the threshold denotes a true claim and rejected if below, again highlighted in equation 2.1. Further discussion on trial outcomes is provided for the convenience of the reader in the appendix A.3.

It can be assumed that the classification system may only have specific knowledge of the concerned speaker from the test and train utterances for a given trial. Other knowledge utilised is assumed not to contain information of the speakers from the trial[1].

To display comparable performance among systems, a standard database is required. The NIST (National Institute for Standards and Technology) evaluations are used to give an interpretation of system performance in an application scenario under different conditions. NIST has provided a very successful vehicle for assessment of telephony speech, though limited by practicable and financial constraints of database collection. Variations of the system used to produce these results have been entered into the NIST evaluations of 2004, 2005 and 2006 by myself along with my colleague Benoit Fauve. This competitive concept using a standard database for speaker recognition has been in place for over a decade. This allows for comparative research of verification systems from across the world. For 10sec and 1conv task conditions in the NIST evaluation we are expected

[1]This is true unless a system with continuous model adaptation is used with true outcomes

to get approximately 30% and 16% respectively with the system used in this thesis. In contrast, advancements during this work have brought performances to approximately 15-20% and 2-4% respectively. These approximate performance values are taken from systems submitted to the NIST 2006 evaluation. NIST evaluations are constrained to assess technology with apriori information which is assured to be correct, known to be determined and labelled automatically. We fall into the trap of assuming such prior knowledge. Throughout this thesis, no errors are assumed to be contaminating the prior knowledge. Further analysis of NIST labelling, with primary concentration in the 2006 database is discussed towards the end of this thesis.

2.2 Classification Framework

The SV system used in this thesis has three primary stages, separated by:

1. Feature extraction of acquired utterances
2. Model generation
3. Scoring/assessment

Depicted in Figure 2.1, this general decomposition can be applied to a large number of classifiers. The increasingly popular support vector machines (SVM) and traditional Gaussian mixture models (GMM) are two such examples. Here, for example, hypothesised scores for verification trials are gathered using the GMM classifier. At the initial stage, feature extraction attempts to distil speaker

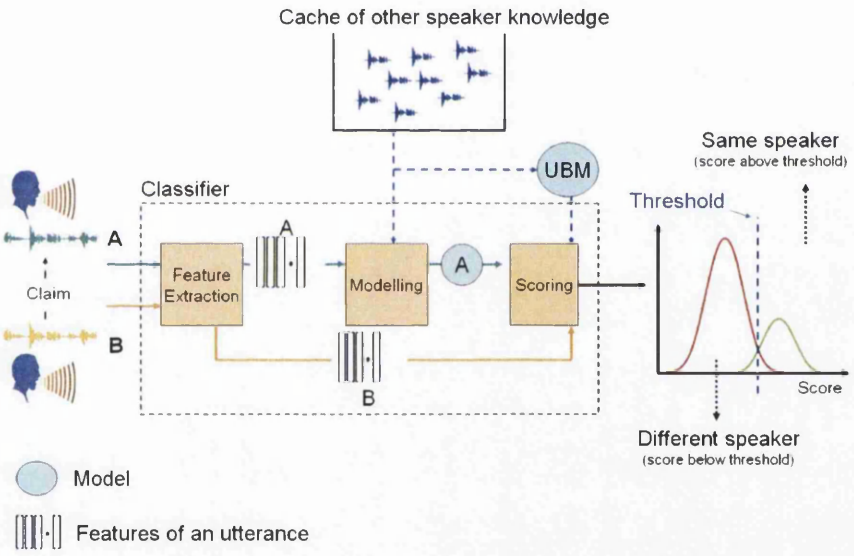


Figure 2.1: High-level overview of a speaker verification classifier showing the stimulus of data and a generalised modular illustration of the processing stages

related parameters from one or more of the supplied speech utterance's. The speaker dependent information can be represented as a set of parameters commonly known as the *features*, extracted using the same parameters for both the claimant and enrolled speaker. For the GMM, features from a sample of speech are quantized into a vector of $1 * N$, where N is the n^{th} order feature coefficient of the speech cepstra. A single vector constitutes a single window of speech, typically 20-30 ms in duration. By sliding the window along the speech utterance, a series of vectors are extracted using spectra derived coefficients by a pre-defined filter bank; further transformed into the cepstra domain to reduce correlation between coefficients. A speech sample is now represented

as a chronological sequence of feature vectors. For traditional SV on a frame level basis, it is usual to post-process the generated features by removing the redundant non-speaker information [1], i.e. silence.

Modelling, groups extracted features an utterance into a collection of probability density functions (PDF's) . This process is also known as enrolment. Generally, an initial model is used to represent a background of speakers, commonly refereed to as the universal background model (UBM) [2]. The UBM is described as collection of multi-variate Gaussian probability density functions, represented as a collection of means, variances and weights. These PDF's are constructed from a large data set of other speech utterances. Each Gaussian component has a dimensional order equal to the number of feature components extracted at the initial stage. This approach employs a cache of speaker knowledge, highlighted as the blue speech icons in Figure 2.1, facilitating observations for the alternative hypotheses as a speaker-independent model from the perspective of the 'world of speech'. The UBM performs several roles, initially it is used for robust model adaptation to compensate for lack of speaker-specific speech, especially for short-duration tasks. Compression of a large collection of utterances is also achieved in the text-independent scenario. In recent publications [15, 16, 17] the UBM has been used as an acoustic reference space to project speakers in this space for subsequent discrimination.

Enrolment is conducted by adapting the UBM using the speaker-specific feature vectors, to produce a speaker-specific model (training of the target model A), also referred to as a target model. A speaker model can be created independently of a background model, using the same approach for generating a UBM, though adaptation of a background model shows better performance [18] . The approach of modelling by adaptation used here is called Maximum A Posteriori (MAP) [18, 19] introduced through the HMM for speech recognition. This procedure updates the well-trained parameters of the UBM, resulting in a speaker-specific model.

We have discussed the method of background model generation and speaker adaptation, the third stage of SV is to verify an unknown test utterance with its claim to an enrolled speaker (our target model). For this process, we have a UBM, created from a large collection of different speakers and a target model for an enrolled individual. To check the validity of a claim the feature vectors of the test speech signal is compared to both UBM (denoting H_1) and the speaker model (H_0). Utterance B is scored against the target model λ_A and results in an observation of B given λ_A, $P(B|\lambda_A)$ also deemed a similarity score Details of the scoring process can be found in [19]. As previously mentioned, the second role of the UBM provides the alternative hypothesis score, $P(B|\lambda_{UBM})$ through the same scoring method. An LLR can then be computed using both these observations, subsequently compared to a predefined threshold θ.

[2] Universal background model is interchangeable with background model, anti-speaker [4], world model [10] or general model [14]

Conventionally, a single policy is undertaken when scoring a trial by creating a model of the training utterance λ_A and examining the generated model against the test utterance B, resulting in $P(B|\lambda_A)$. The converse of this procedure is the *bilateral*. Here, a model is created using the test utterance λ_B for which its model is examined against the training utterance A, to give $P(A|\lambda_B)$. Preliminary examinations have been conducted on the fusion of score outputs from both policies or dismissal of the recessive score, providing greater performance increase across different evaluations. This thesis only considers the single perspective for analytical clarity, though the bilateral approach can be applied to all experimental results discussed, usually with greater robustness.

The closest component in the UBM to which the test feature vector matches is used as an accumulated *similarity metric* for H_1. As UBM adapted speaker models are used here, the same PDF component in the UBM is compared in the target model. If no adaptation has taken place for the same component in the speaker model, the differential score is zero. If the component has been adapted, a similarity measure is given. The higher component deviation from the UBM, the greater the similarity score. Similarity metrics can be swayed when degradation is embedded in the utterance, causing features and in-turn models to be susceptible, enhancing unwanted variation in the test stage on the similarity metrics, making scoring less robust. These distortion can be reduced using a variety of normalisation approaches.

2.3 Normalisation

The assistance of additional sources of speaker knowledge can improve verification performance to help overcome the need of speaker-dependent decision thresholds with a single evaluation specific threshold. The use of prior knowledge[3] in text-independent SV is commonly used through many normalisation approaches in verification systems. For example, in the Gaussian mixture model (GMM) framework, some lack-of-speech properties are reduced by enrolling a specific speaker through an adaptation process from a pre-generated, well trained representation of the world, e.g. the UBM. Enrolled speakers are represented as models of which are projections of usually only the training utterance. A speaker-independent model assists in several ways, aligning text-independent speaker utterances via adaptation, utterance compression and more recently exploited as an acoustic reference space for subsequent discrimination through an SVM [21]. For the text-independent scenario, the time-sequence information is also reduced[4]. However, Stapert [22] attempts to retain time-sequence information by fusing such characteristics into the speaker models.

[3]Prior knowledge can also be interchangeable with *auxiliary information* [20]

[4]A form of short-term time-sequence information can be included through the delta feature approach and its derivatives

Prior knowledge from other speakers and their speech can enhance the background data set to assist with the degenerative variance between the two supplied utterances for a trial by removing unwanted variables. Gender, for example, is a simple binary filter, utilizing prior knowledge; if the test utterance was spoken by the opposite gender to the claimed target, it can be quickly rejected. For a given trial, the umbrella of prior knowledge can be broken down into three sub-categories:

1. **Primary**: The finite amount of speaker-specific utterances supplied for the trial (an assumption is made that no further model adaptation is applied after a successful verification). This includes the correct labelling of the primary utterances.

2. **Secondary**: Practicable utterance specific information, e.g. gender and microphone, supplied with or derived in an online manner from the two utterances under examination. A sub-set of such pre-labelled knowledge is assumed to be correct and provided with the NIST evaluation utterances.

3. **Other**: Additional material from other sources prior to the considered trial. Our knowledge of speech.

The question we pose is, how do we make best use of this knowledge to compensate for the discussed variability's to aid classification?

Ferrer et al. [20] recently published the use of prior knowledge, here called *auxiliary information* by combining such information as length of utterances, gender, nativeness, etc. of a given speaker through a combinatorial mechanism to aid the decision making. Here a data-driven approach is utilised to estimate a weighting regime for score level fusion of the speakers extra knowledge. Published results showed that for the NIST evaluations, native or non-native English speaker scoring gave good enhancement for combination with other classifiers.

A variety of established normalisation methods endeavour to compensate for unwanted signal degradation between two utterances under trial, enhancing the speaker-specific content. During this thesis, discussed approaches for unwanted signal compensation can use combinations of prior knowledge from the three derived sub-categories.

Unwanted non-speaker signals are preserved from the recorded utterance through to all stages of the classifier with unwanted preservation in the computed score observation. With the large variation among and during recording sessions due to the nature of the changing voice and other degradation effects, the need to normalise out these session variability's is just. Much research has been conducted at different stages of the SV system in an attempt to suppress/normalise forms of non-speaker descriptive signals.

Filtering methods can be used at the signal level during feature extraction to reduce both slow and fast varying convoluted signals from the utterance. Cepstral Mean Subtraction (CMS) and RASTA [23] attempt to reduce these convoluted effects. CMS is a simple form of normalisation is to remove the d.c. line component from a recording session. Derivatives of the cepstra, denoted as *delta* and its second derivative *double-delta* may also be applied to the features to account for short time-base transitional events along the chronological axis of the feature vectors. However, some of these extra features can add perturbations onto the features for different conditions, shown by Fauve et al [24]. Kuhn et al. [25] generates eigenvoice based vectors to model variation between different speakers, ported from the field of face recognition. A data-driven method to obtain orthogonal basis vectors which represent the most important components of inter-speaker variation between a development set of speakers. A linear combination of voice components from a variety of speakers where each component obtained through principal components analysis transformation where eigenvoice 0 represents the mean component. Removing a number of the upper principal components provides model compression. Kenny et al. applies this eigenvoice vector to normalise the target models via a common reference position and assists with sparse training data [26].

Nuisance attribute projection (NAP) [15] and factor analysis [26] are two examples that apply normalisation to the modelling stage of the now popular GMM-SVM hybrid system. Models generated with the addition of either approach are now normalised models and treated as input features to a subsequent classifier, the SVM. This second classifier discriminates models (the normalised model is also referred as a GMM supervector) in the normalised acoustic domain with *normalised features.* This two-stage classifier known as GMM supervector linear kernel SVM-GSL [24, 17] shows large performance gains on long-task durations. Fauve et al. [24] has shown negative performance for NAP on short-duration task evaluations due to the nature of their session variability and unpredictable phonetic content. Short utterance variability is also acknowledged by Li and Porter [27]. NAP weights the feature components by their relevance of contribution to the speaker information within the speech signal. Low weights are given to the nuisance feature coefficients. NAP is an example of performing degenerative signal compensation at the modelling phase for discrimination through a two-tier GMM and SVM (SVM-GSL) classification system. An alternative is factor analysis, similarly to NAP, this method applies noise compensation to the speaker models. It is postulated that each speaker-specific model is an additive of both speaker and channel related information and can hence be separable. The state-of-the-art trend that has developed during the work presented here appears to move the concentration of utterance normalisation to the modelling or speaker enrolment domain. In this thesis however, the investigation into impostor-centric score normalisation concentrates on selection approaches such as the state-of-the-art adaptive T-Norm. Such approaches discussed here for use in score normalisation could possibly be adapted to assist with such data-driven approaches as NAP.

Speaker model synthesis [28] and Feature mapping [28] are other methods applied at the modelling

level to reduce, for example, channel variability by building channel dependent models and mapping the Gaussian components from the UBM to that of the specific channel model during the scoring routine. This is similar to a collection of speaker-independent models used in cohort normalisation in section 2.3.

For conventional GMM scoring, the input test utterance for a given trial has not been normalised in contrast to the target model. The SVM-GSL accounts for this by discriminating both the training and test models in the model domain. Differentiation on a normalised model-to-model basis alleviates much unwanted signal perturbation embedded on the test utterance. Normalisation for both test and train is not possible with a traditional GMM classifier approach[5]. Different sources of prior information can be modelled in different systems for each target speaker and then combined at the score level through some weighting function during a trial.

Score normalisation is the final form of compensation applied to a classifier prior to making a decision. Zero normalisation (Z-Norm), handset normalisation (H-Norm) and test normalisation (T-Norm) with subsequent extensions have been used to enhance the robustness of the decision threshold by reducing the variability of resulting true and false distributions. Normalising the scores, emitted from a classification system can assist with these variations. Applying a score normalisation method is made prior to making a decision is made.
Two score distributions are the produce of an evaluation. A development set has coupled with it a ground-truth for each conducted trial. These distributions display the separable nature of the evaluation as binary outcomes of true, target claims and false, impostor claims. These methods attempt to bias and scale these distributions to enhance performance. Discussions by [11] state that research has focused on adjusting the impostor based distribution, with a side-effect of adjusting the true distribution simultaneously. This is a practicable, application driven constraint as the availability of target speech can be a rarity, though an abundance of impostor speaker models can be available to any given trial. This is an impostor-centric approach.
Score normalisation but primarily test normalisation (T-Norm) and its derivatives applied on the GMM classifier is the main focus of this thesis.
The methods discussed to reduce unwanted signals at the model and scoring stage use *other knowledge* category of available speech (described in Chapter 1.1) to derive normalisation statistics. Although many of these techniques can be applied at various stages throughout the classification system, it is difficult to predict the contribution of normalisation to downstream from methods such as score normalisation when techniques further upstream, such as NAP are applied. It would be interesting to observe the contribution of different normalisation approaches when combinations are configured, highlighting the usefulness of such methods. It is also important to note that these methods can be applied independently at all stages with varying effectiveness on the reduction

[5]Unless performing direct model-to-model comparison, for which the SVM approach contributes

of signal degradation. An example of concatenating normalisation approaches are presented by Barras and Gauvain [29] where it was found that assistance of both feature warping and T-Norm gave enhanced robustness.

2.4 Score Normalisation

From herein, the traditional GMM framework for SV will be the primary focus for modelling and scoring and used as a template for describing the stages of score normalisation and application to subsequent experiments.

The hypothesis of factor analysis shows that the speech signal is a combination of speaker and channel related information that can be separated into two vectors. This is wholly analogous to other normalisation techniques by generalising on the channel related information to other forms of non-speaker signals.

Based on the discussion of hypothesis testing in section 2.1, the given model for H_0 is well defined and can be easily estimated from the given claimed speaker-specific speech. However, hypothesis H_1 has no specific criteria to construct an alternative model other than any information that originate from either trialled utterances. A landmark for SV was the introduction of the speaker-independent background model, popularly referred to as the universal background model (UBM) was introduced by Carey et al. [14] as the *general model* in the hidden Markov model (HMM) classifier to obviate the need for an absolute threshold. Gravier and Chollet show that the stability of a threshold when applying either Z-Norm or H-Norm is greater [30].

Preceding the institution of score normalisation for SV, the impostor cohort normalisation [31, 32] approach substituted the UBM as the alternative hypothesis (although the UBM is still used to provide the template model for speaker adaptation). A cohort of N UBM adapted speaker-dependent models replace the single speaker-independent model that encompasses the world of speakers. Equation 2.4 illustrates that some function $f(.)$ can represent a statistic, e.g. maximum, average, etc., to represent an alternative hypothesis score from the N available speaker-dependent models.

$$P(B|\lambda_{\overline{hyp_1}}) = \{f(P(B|\lambda_{I_1}), P(B|\lambda_{I_2}), \ldots, P(B|\lambda_{I_N})\}) \tag{2.4}$$

Cohort normalisation with a tailored speaker set deemed to provide better performance over the traditional UBM similarity normalisation [31, 19]. It is interesting to note that if the cohort contained the target speaker, given a large N, the average of the cohort would absorb the influence of the target. Strictly speaking, the speaker under test (the target) should not be represented as one of the cohort models as this would violate the hypothesis H_1. Applying a maximum statistic criteria

would surly violate the alternative hypothesis if the target speaker was chosen as the maximum score.

It has been shown that scores from multiple UBM's can be used to represent the alternative hypothesis; again tailored and trained to specific application scenarios. Gender is a category that could be exploited by generating independent male and female UBM's for subsequent speaker-specific adaptation and scoring. Secondary type prior knowledge (defined in Chapter 1.1) constitute different channels, genders or other practicable information that can be applied.
Applicably, when conversing over a mobile network, each person could likely have different handsets, displaying miss-matched signal perturbations. These ambient characteristics can provide distortions, impeding speaker discriminative information. A process to reduce such effects is the use of handset normalisation (H-Norm) [7, 19, 33], even after applying several standard linear channel compensation approaches (e.g. CMS and RASTA).
H-norm is one method from a family of normalisation approaches that operate in the score domain; applied at the final stage of classification prior to making a thresholded decision; attempting to reduce unwanted signals between the two utterances of a trial.

Score normalisation is key to performance, establishing a shared decision threshold between speaker trials by removing the influence of variability between utterances. An enhanced discriminative threshold can be found by altering both true and false distributions for an evaluation by altering the individual trial scores. This is achieved by modifying individual hypothesised scores to a common scale, helping to reduce the effects of mismatch between training and test utterances. The basis of general score normalisation is to centre the impostor distribution by applying equation 2.5 on each trial.

$$S(\lambda_A|B) = \frac{\Lambda(B) - \mu_{\lambda \subset I}}{\sigma_{\lambda \subset I}} \tag{2.5}$$

Where μ denotes the statistical mean of scores derived from a set of speaker models from cohort I. σ represents the statistical standard deviation from a set of speaker models I with $\Lambda(B)$ (equation 2.3) dictates the log likelihood ratio of a trial. The resulting true and false score distributions are simultaneously scaled and biased when applying equation 2.5 to each trial of an evaluation. This is an *impostor-centric* approach to score normalisation, originally introduced to the SV domain by Li and Porter [27]. It is a form of distribution scaling where one of the distributions are aligned. Aligning the false distribution through utilising the impostor models by score normalisation is considered as impostor-centric scaling [13]. Large variances were observed by Li and Porter for both impostor and target scores and hence reduction of such variation by focusing on normalising these scores for a set of trials. The impostor models are generated from a development corpus, different to the corpus of evaluation. This approach was applied in 1997 by Finan et al. [32] in the text-dependent SV mode to set a single threshold for all speakers.

Applications such as speaker diarization can produce an abundance of speaker-specific speech with possible normalisation through a target-centric approach. Such scaling is the complementary process to impostor-centric normalisation, aligning the true distribution of the two-class problem. Generating normalising statistics by using an arrangement of models generated by the same speaker of the target. Utilising the normalisation based on an impostor based cohort is guided by two factors, highlighted in [1]. First, psudeo-impostors are more readily available in most application scenarios where target-specific based cohorts would be difficult to collect through lack of target speech. Auckenthaler [13] describes the instance of the ever changing voice, especially with illness can effect the target score, causing high false rejection rates in the target-centric case. Speech verification within a scenario where a system can continuously utilise an abundance of target specific information would reduce this effect somewhat. Secondly, the false score distribution for an evaluation represents the largest deviation. Li and Porter demonstrated this when scoring speakers at the segment level, with a train of scores being the resultant of an utterance. They observed that the variance of impostor scores over segments varies widely and this variability could be help stabilise, i.e. normalise the parallel target scores on a segmental level. They also state that the accumulation of scores becomes optimal if the distribution becomes Gaussian.

The score normalisation procedures discussed in this thesis can be applied to any classifier (e.g. HMM, GMM, SVM) that produces a score based on some observation. These are primarily data-driven techniques, widely used in established normalisation procedures such as Z-Norm and T-Norm. Sources with groups of variability, for example, utterances with common duration signatures, can be used to reduce score miss-match between both training and test utterances.

Popular normalisation procedures published to carry out score normalisation will be discussed in the following sections.

2.5 Z-Norm

Zero-Normalisation (Z-Norm), derived from Li and Porter observations [27] attempts to align between-speaker differences by producing statistical parameters for each speaker-specific model, using a cohort of widely available impostor utterances to align the impostor scores to zero. For a cohort of E impostor utterances, are scored on each enrolled speaker model, e.g. λ_A to generate a series of likelihoods, shown in equation 2.6.

$$\{P(\lambda_A|I_1), P(\lambda_A|I_2), \ldots, P(\lambda_A|I_E)\} \tag{2.6}$$

This is a speaker dependent process where the mean and standard deviation statistics are extracted from this impostor distribution of scores which are fixed for all subsequent trials conducted on the speaker-specific model λ_A. These model normalisation statistics can be generated off-line during speaker training for each speaker and are applied for each trial using equation 2.7.

$$S_{Z-Norm} = \frac{\Lambda(B) - \mu_{impostors}}{\sigma_{impostors}} \tag{2.7}$$

2.6 H-Norm

Using a variety of capture devices to record a speaker over different sessions can yield convoluted non-speaker signal characteristics between utterances. An issue raised by attributes of different handsets or microphones. Utilizing secondary type prior knowledge, handset normalisation (H-Norm) attempts to remove speaker-independent d.c. offsets from the target scores. Impostor utterances from different labelled handset types are used on an adapted speaker model to generate handset-dependent constant scaling parameters, i.e. mean and standard deviation per handset type for normalisation. Statistics for an electret and carbon button handset type for a specific speaker model is shown in equation 2.8. Similar to Z-Norm, normalisation statistics are calculated 'off-line'.

$$\{\mu(CARB), \sigma(CARB), \mu(ELEC), \sigma(ELEC)\} \tag{2.8}$$

Analogous to Z-Norm, the statistics are traditionally applied on a likelihood score with equation 2.9. The handset type is supplied as a secondary type prior knowledge which can be used to select the appropriate normalisation statistics for a trial. An example of carbon button handset normalisation is shown with equation 2.9. The carbon button statistics can be replaced with statistics of the appropriate handset. As with Z-Norm, each target model is supplied with a personal set of H-Norm statistics for each considered handset.

$$S_{H-Norm} = \frac{S_{target} - \mu(CARB)}{\sigma(CARB)} \tag{2.9}$$

Wu et al. has generated impostor cohorts by matching emotional states of target speakers. Similarly to H-Norm, termed 'Emotion-dependent score normalisation' or E-Norm, this approach has proven effective [34] in reducing miss-match from the two trial utterances.

2.7 D-Norm

The distance normalisation [35, 36] (D-Norm) approach determines a normalisation statistic at the scoring stage. This approach does not employ additional data and hence does not require the need

of utterances as stimuli to generate distributions from where, for example T-Norm gathers normalisation statistics. The D-Norm could advantageously replace data-driven approaches in some real-world applications scenarios when little or no additional data is available. In contrast, for a data-driven approach, the greater the amount of training data for a speaker, the larger the distance on average are the impostor scores generated from this model and hence less similarity (a greater score). By applying a model-to-model differential entropy distance calculation, the 'distance' between a target and background model can be applied to normalise the score distribution. This becomes important when applications gather utterances of different lengths over evaluations where different thresholds are required between tasks of different training duration; compensated by applying this normalisation approach. Although reports show that small gain is achieved, in the NIST domain, this is not as crucial as the evaluation tasks are primarily separated into categories by their training and test utterance duration. However as discussed in later chapters, some subsets of training utterances contain unexpected duration's, for which the application of the D-Norm approach could naturally overcome. The results reported by Ben et al. [35] show comparative, though slightly worse results to the Z-Norm approach. This process is not examined in detail in this thesis as we have an abundance of impostor utterances to generate normalisation statistics using data-driven approaches which are reported to give better performance.

2.8 T-Norm

In the NIST evaluations of 2006, approximately 58% of participating institutions applied the popular test-normalisation(T-Norm) approach to their systems in one form or another. T-Norm [10] was proposed to easily set a speaker-independent threshold by reducing the effect between test utterance scores over different trials. This is achieved through scaling and biasing the true and false score distributions over an evaluation.

An impostor score is defined as a LLR ratio by equation 2.10, similarly to the LLR of a target, illustrated previously by equation 2.3. Here, a cohort of impostor models of N quantity are used resulting in N scores to derive statistics for normalisation, highlighted in equation 2.11. A similar process to Z-Norm though focusing on the test utterance instead of the training utterance. In contrast to Z-Norm, the test normalisation statistics are computed in an 'online' manner at test time for each trial.

$$\Lambda(B)_{I_N} = \log(P(B|\lambda_{I_N})) - \log(P(B|\overline{\lambda_{UBM}})) \tag{2.10}$$

$$\Lambda(B)_{I_{1\ldots N}} = \{\Lambda(B)_{I_1}, \Lambda(B)_{I_2} \ldots \Lambda(B)_{I_N}\} \tag{2.11}$$

$$S_{T-Norm} = \frac{\Lambda(B)_T - \mu\{\Lambda(B)_{I_{1...N}}\}}{\sigma\{\Lambda(B)_{I_{1...N}}\}} \tag{2.12}$$

During UBM normalisation, the UBM is used to bias the test score for a target model. The influence of the biasing UBM score during T-Norm is cancelled as the bias from both target likelihood and cohort mean include the UBM scores. Proof of this is explained by Navrátil et al [11]. With the UBM influence reduced (though its influence from adaptation is still required to compensate for lack of speaker-specific speech), the biasing and scaling results from only the N impostors observations are present.

An interesting hypothetical scenario discussed by Navrátil [11] depicts when the T-Norm method becomes redundant if both true and false distributions are identical and Gaussian. In real-world applications, this scenario would be unlikely. Similarly to the observation that H-Norm is to Z-Norm, HT-Norm is a variation of T-Norm where impostor models are selected by a criteria of handset type from where the speech was captured. Normalisation statistics are produced using the same approach to T-Norm with chosen impostors being specifically matched to the attribute of the speaker handset used. Concatenation of score normalisation can also be applied, examples of this are the use of both Z-Norm and T-Norm, resulting in ZT-Norm or TZ-Norm [37]. The order of 'Z' or 'T' reflects the sequence of the conducted normalisation approaches. ZT-norm is reported to give optimum results with approximately 3% enhancement over standard T-Norm. Utterance level T-Norm is the direct equivalent of T-Norm in the text-dependent scenario. Recently, Toledano et al [38] introduced the phoneme-level T-Norm to the text-dependent scenario and later discussed the state-level T-Norm. This is analogous to normalising numerous segments of feature vectors along an utterance in the text-independent scenario.

A Cohort for T-Norm

The duration of an utterance used to train a model can vary, reflecting the nature of an adapted model. Valid for both target models and impostor models that represent the cohort. The higher duration of speech given for training, the greater potential of UBM adaptation to a specific-speaker model. Considering this variation between models, the distance between model components of a target and impostor can be observed loosely as a *quality measure*. Models trained with more speech duration tend to give a higher overall *similarity measure* during a trial, where the components in the model have received greater adaptation. The scores generated by short utterance trained speakers have low scores between speakers and high scores with well trained speakers. Figure 2.2 demonstrates this with the use of one sample test utterance, common to a pool containing 600 T-Norm impostor models. The similarity scores for 600 individual T-Norm impostor

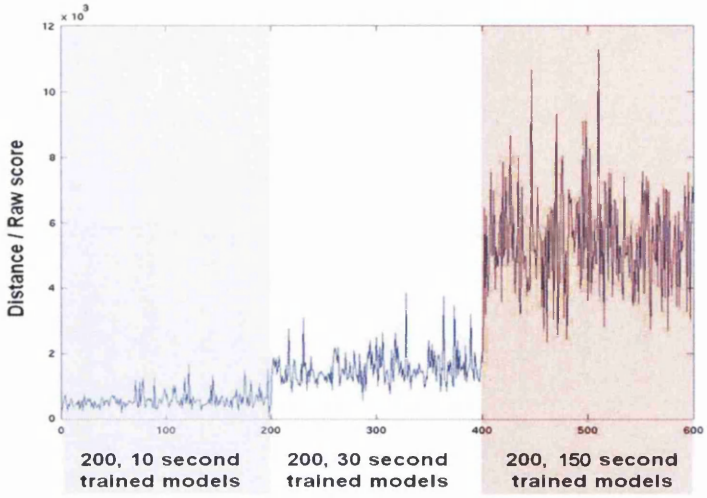


Figure 2.2: The x-axis resembles a group of 600 speakers, divided into three equal groups of 200 speaker models, highlighted in groups from the 10sec, 30sec and 1conv conditions. The increase of training utterance length is from left to right on the x-axis. The similarity scores from trials of these speakers against the same test utterance is shown in the y-axis, illustrating the variability of the score metric prior to score normalisation (or world normalisation)

models are speaker models trained with three different utterance durations of approximately 10, 30 and 150 seconds (1conv) of speech shown as sequential blocks of 200 models along the x-axis respectively. The scores are generated from an independent set of test utterances to generate scores for all 600 T-Norm models. The similarity scores generated increase with models of higher speech adaptation, on average showing an increasing trend with more training material used to generate a speaker/impostor model. This is similar to the reported results by Pelecanos et al. [39] where the variance of impostor scores increases when more training data is applied to the models. The model will eventually saturate with enough speaker-dependent data, however, utterances from the shorter duration tasks are primarily considered in this thesis. This highlights the idea proposed in [11, 40] to select impostor models that are representative (based on the duration of speech used for speaker training) of the hypothesised model, empirically found to show enhanced robustness with short-duration tasks (i.e. a 10sec test).

Selecting the impostors to generate normalisation statistics is an intriguing attribute of the T-Norm. As an example of scores generated by a cohort, Figure 2.3 shows the variability of a fixed T-Norm cohort with 10 different test utterances, identified along the x-axis. In Figure 2.3, we can observe that for a fixed cohort of the same 600 models as previously described (depicted by the red scatter crosses), T-Norm presents a large diversity of impostor-centric scores given different test utterances. Changes for both mean and standard deviation can be observed. Test utterance 3 is of particular interest. Demonstrating a low variety of scores for all impostors with the particular

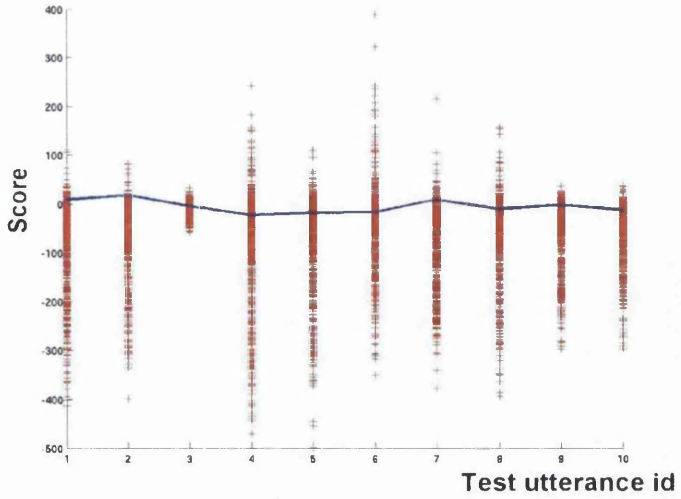


Figure 2.3: LLR scores using the same target model against a multitude of test utterances (x-axis). The same 600 T-Norm models are used for each test utterance, with scores (y-axis) depicted by red crosses. All trials are assumed impostors.

test utterance. It was observed that the contents of the test utterance contained little speaker-discriminate information, with a few repeated gestures of *"mmmmm"*. From the assumption of impostors utterances being tested against impostor models, a resulting high score would result in an error. This model would likely give false acceptances if used in a real trial. As we assume that these tests should generate a false, low score, then these high scores are dubious. Figure 2.4 shows 50 detection error trade-off (DET) [41] curves of the 10sec-10sec evaluation from NIST 2005 by selecting different impostor cohorts from a pool containing 600 male and female impostors. Subsets comprise of 200 *10 second*, *30 second* and *2.5 minutes* of utterance duration from the NIST 2004 evaluation database. We can observe the variability of performance in Figure 2.4 when selecting in impostor cohort can have a detrimental effect. Different training and test conditions, including 1conv4w-1conv4w will also be investigated in subsequent chapters.

The motivation of this thesis originates from the intriguing degree of sensitivity presented in Figure 2.4, utilising different compositions of impostors to represent a cohort through the T-Norm approach. An investigation of this phenomena will now be illustrated empirically and its ramifications discussed when specific impostors are not carefully selected.

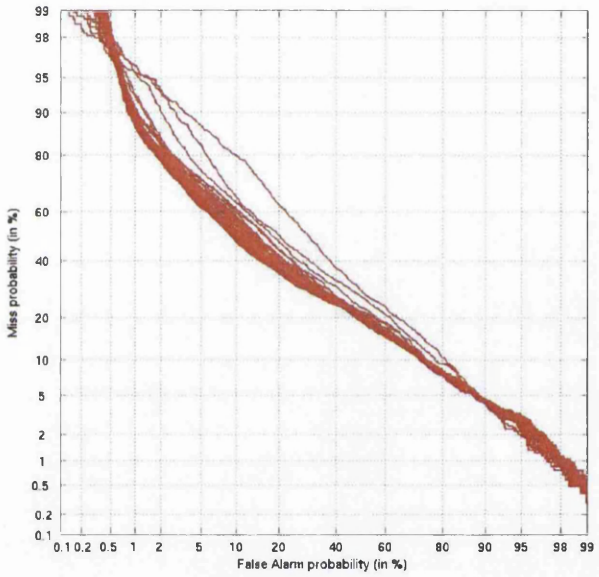


Figure 2.4: This is the general configuration of the DET plot where the x-axis represents the False Alarm probability and the miss probability is represented by the y-axis and used for all forthcoming DET illustrations. The profile shows 50 DET plots for 50 different selected impostors from a mixed pool

Chapter 3

Trial-independent Cohorts for T-Norm

Reynolds [19] informs of two issues that arise with the use of background speakers, a selection regime and its quantity. In this chapter we follow these observations with score domain normalisation, in particular using the T-Norm approach, discussed in the previous chapter. For the impostor-centric approach, Reynolds suggests that the ideal number of background speakers should be as large as possible to better model the imposter population, but practical considerations of computation and storage dictate a small set of background speakers. This is an implementation criterion [19], the more similar the models, the smaller the ratio of required models becomes. Similarly, Doddington et al. also states that the impostor population should model the target population [3]. This chapter examines the theme of specific cohort selection of impostor models for test-normalisation over all trials of an evaluation. Here, denoted as *trial-independent* selection. *Trial-dependent* cohorts will be discussed in Chapter 4.

3.1 Introduction

The foundation of T-Norm has been established for speaker verification (SV) in Chapter 2. Here we shall investigate selection procedures of impostors to compose a cohort.

Primarily, we can ask, what prior knowledge can we exploit to address selection? For example, Auckenthaler et al. [10] shows that this can be achieved by discriminating between utterances of different durations are easily obtainable, gender is another. Conventional T-Norm sets to use some broad speaker specific criteria to perform selection [19]. Gravier et al. [4] shows the application of prior-knowledge through handset information into the classification system through the z-norm normalisation approach, with either handset or gender information. In the NIST evaluations, gender, handset type, coding characteristics are other sources of prior information are available.

We can denote two umbrellas for cohort selection, *trial-independent* and *Trial-dependent.* Trial-independent (or target speaker-independent) selection is the conventional T-Norm strategy, using

a fixed set of impostors for all trials over an evaluation condition. Experimental results rendered in section 2.3 used a trial-independent approach to perform normalisation. Trial-specific selection supplies an independent set of cohorts to each speaker model, introduced by Sturim & Reynolds as the adaptive T-Norm (AT-Norm) [12], discussed further in Chapter 4. Reynolds [19] postulated that two issues that arise with the use of background speakers, the amount of speakers to use and its composition. When selecting a subset of models to obtain normalisation statistics, computational efficiency is a constraint along with the availability of a realistic and finite cohort. Varying this finite number leads to a selection of a subset of E impostors. Selective criteria including, random & general set, matched and its converse miss-matched are denoted as trial-independent approaches whilst trial-dependent strategies constitute adaptive selection (data-driven) or model distance (model to model) selection. The goal of a trial-dependent selective T-Norm strategy is to identify a set of impostor speakers that represent similar characteristics of a target speaker. These conditions can, for example, be the similarity between speakers based the utterance duration.
Ideally the number of background speakers should be as large as possible to better approximate the background population, but practical considerations of computation and storage dictate a smaller set of background speakers. This is an implementation criterion.

Impostor-centric normalisation coined in [13] is used to align the scores of the impostor distribution, which will be the focus here to generate the normalisation statistics. As discussed in section 2.4, the cohort composition with the inclusion of the target model would violate the LLR hypothesis H_1. Empirically, this in not true; for a large enough score normalisation cohort as its influence would be drowned among other models.

Apriori selection can be done from knowledge of qualitative measures, e.g. gender, age, environmental noise, transmission channel/coding or quantitative duration of utterances (or number of target-specific utterances). Variation in the test and training utterances constitute between speaker differences, e.g. utterance length. Speaker differences can also occur between different recording sessions. This can be reduced in applications over different sessions. After a successful verification has occurred the target model can be further adapted, becoming more representative of the particular speaker. This can assist the adaptation of a persons voice through age and of course collect more speaker specific data to enhance their models. Further adaptation is illustrated in Figure 3.1 by increasing the availability of speaker specific speech. For mean-only MAP adaptation, more components can be influenced for potentially greater discrimination. Beginning with the UBM in Figure 3.1, this simply illustrates the diversity a speaker model when the UBM mean components are adapted with utterance durations of 10 seconds, 30 seconds and 2.5 minutes. For a target and impostor cohort, it seems logical to use models that produce similar observational score statistics. Conceptually comparing *like with like.* Consider a large pool P of 600 impostor models for T-Norm, divided into 3 groups of approximate duration; 10 second speech duration for a speaker (10sec), 30

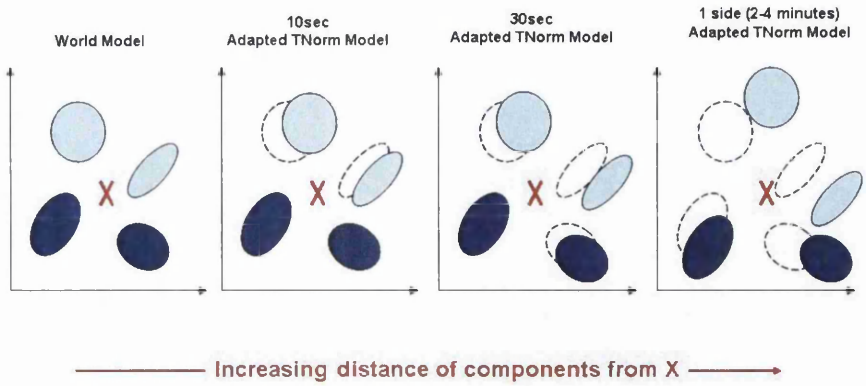


Figure 3.1: A simplified two-dimensional illustration of a mean only MAP adaptation process when more target-specific speech is utilised. Both axis in each picture is an arbitrary feature dimension.

seconds of speech (30sec) and 2.5 minutes of speech (1conv). During the discussion of experimental results, colours are representative of the speech duration of a speaker, red for 10sec impostors, green for 30sec and blue for 1conv. Evaluations concerning the whole pool of 600 impostors will be depicted as black. All impostor speaker models from pool P are taken from the NIST 2004 evaluation database. For the duration of the thesis, terminology *matched* and *miss-matched* will be

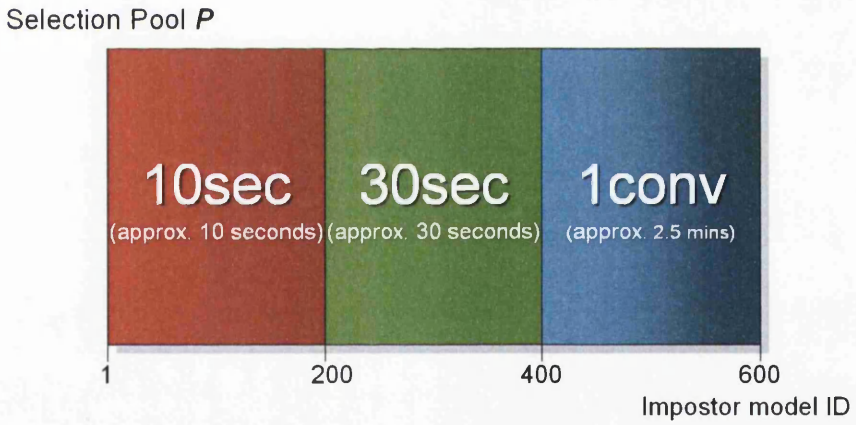


Figure 3.2: The selection pool P containing defined subsets of impostors models based on their approximate training duration, 10 seconds (10sec), 30 seconds (30sec) or 2.5 minutes (1conv/1side) of speech with the all pool scenario consisting of the 600 impostors is represented by the x-axis. The y-axis is unity.

used to describe the relationship between impostor and target models. A *matched* cohort scenario considers characteristics of a selected cohort (e.g. duration) to match that of a target model or evaluation (e.g. 10sec impostor T-Norm models are matched to the 10sec-10sec or 10sec-1conv

NIST evaluations). A *miss-matched* situation arises when the selected cohort has different criteria to that of the training condition of the target and/or evaluation, e.g. utterances of 30sec or greater are miss-matched impostor models to the training duration of a 10sec-10sec or 10sec-1conv NIST evaluation. Similarly, the 1conv models are matched for a 1conv-1conv or 1conv-10sec evaluations. Dunn et al. [42] applied this approach to the H-Norm in 2000 when matched and miss-matched coder characteristics were used to generate the score normalisation statistics. It was found that a fully matched coder cohort gave overall better performance than a miss-matched cohort.

System performance will be illustrated through the detection error trade-off (DET) curve with the aid of two important interpretations. The equal error rate (EER) shows performance when both false alarm and miss probability are equal with respect to the system performance, i.e. no trade-off. The second parameter is the detection cost function (DCF) , defined for the NIST protocol to evaluate a specific application scenario when the ratio of a *'cost for false alarm*(i.e. accepting an impostor) is 10 times greater than the cost of rejecting/missing a target. Further discussion of these attributes can be found in [41] with a simplified discussion in the appendix A. Best performance is depicted by low EER and low DCF. Protocols for development and evaluation with system configuration can be found in the appendix A.

Initially, we shall investigate the composition of the 50 verification curves briefly described in Chapter 2, for further discussion, a further breakdown is illustrated in Figure 3.3. The arrows on Figure 3.3 highlight three performance curves of interest. The same NIST 2005 10sec-10sec evaluation has been undertaken for all 50 outputs, where each plot is T-Normed from a cohort of 50 models, acquired from pool P. Three plots of interest are depicted by C_1, C_2 and C_3 with their cohort composition shown in Table 3.1. C_1 displays best performance from this random selection, containing a majority of matched 10sec models, i.e. a greater influence of a matched cohort. The declining inclusion of matched models within the cohort, yields less performance, shown by C_2 and C_3.

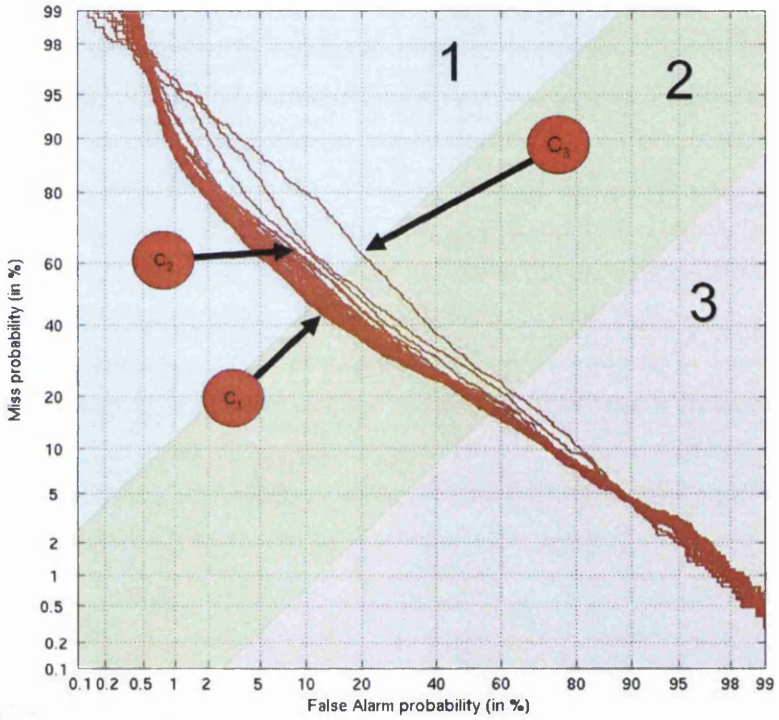


Figure 3.3: 50 DET plots of 50 randomly selected impostors from a mixed pool. Again, this is the general configuration of the DET plot where the x-axis represents the False Alarm probability and the Miss probability is represented by the y-axis.

DET plot id	% of 10sec	% of 30sec	% of 1conv
C_1	87	12	1
C_2	61	9	30
C_3	7	2	91

Table 3.1: Composition of cohort by percentage of models used in three significant performance scenarios for the 10sec-10sec NIST 2005 evaluation using 200 matched (10sec) & miss-matched combinations (30sec & 1conv). C_1 has higher density of 10sec impostors, matched to the training condition of the trial, giving enhanced robustness.

For this 10sec-10sec evaluation a large contribution of high miss-matched seriously degrades performance when T-Norm is applied. This is manifested by the poor performance of C_3 with an EER of 40%. UBM normalisation (not shown here) has 31% EER, greater performance than the T-Norm cohort with high miss-match of selected models. Not all of these profiles shown in Figures 3.3 and 2.4 was collected by randomly selecting impostors. Several impostor cohorts contained intentional assortments, with a majority of either matched or miss-matched were engineered to illustrate the large variability and danger of selection with the 10sec-10sec short duration task.

The highlighted regions on Figure 3.3 show different system operation characteristics, illustrating a range of threshold points, depicted to assist the narrative during experimental analysis. Setting a high score decision threshold in region 1 allows applications with *high security*, e.g. door access to reduce the number of impostor acceptance errors. The equal trade-off or 'equal error rate' (EER) resides in region 2 where an equal probability is given to both target (miss) errors and impostor (false alarm) errors. The third region manipulates thresholds for a *high acceptance* criteria, the converse of high security, e.g. surveillance, where the probability of missing a target is lowered and the acceptance of impostors is increased by setting a lower decision threshold.

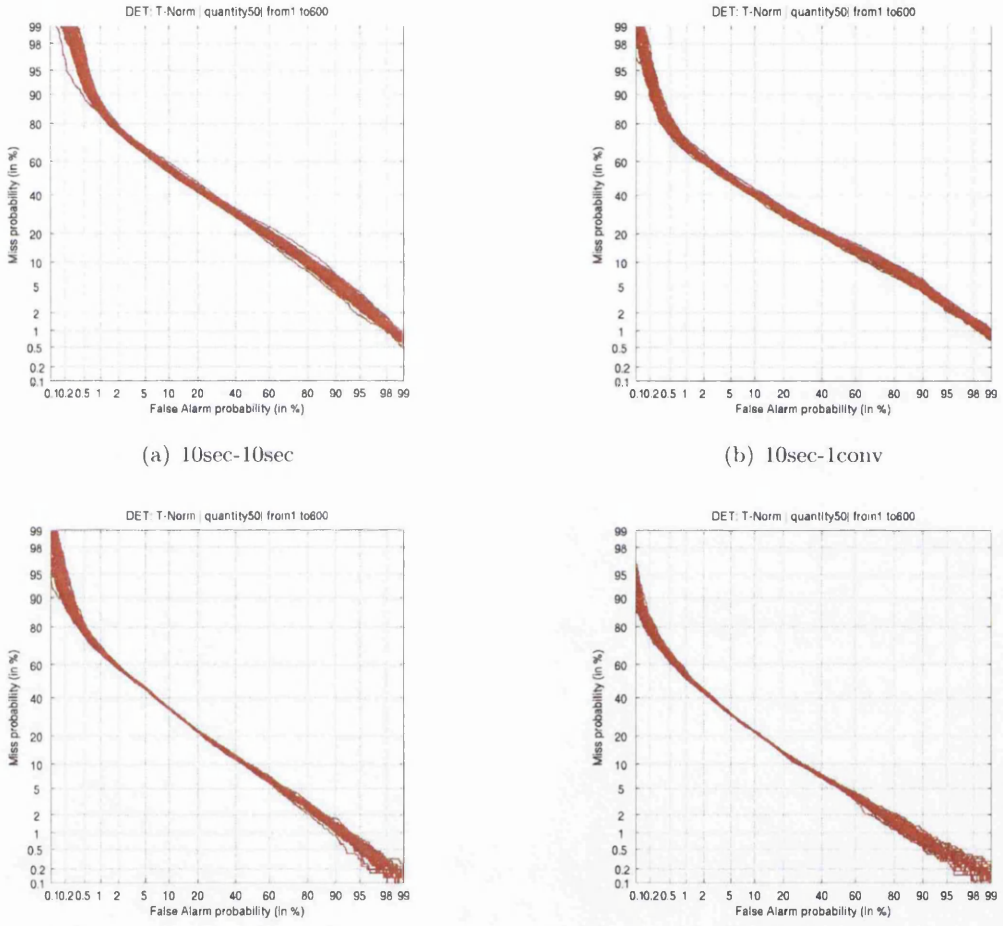
3.2 Experimental Outline

SV experiments outlined in this thesis share many common configuration parameters from a conventional GMM based classifier, detailed in appendix A.1. Primarily we will be concerned with a variety of 10sec and 1conv test and training configurations from the 2005 NIST evaluation. Concentrating on compositions of the T-Norm impostor cohort, the recently described model pool P will be used.

Experimentation will begin with a true random selection of cohorts over different evaluation conditions. The match and miss-match selection by exploitation of utterance duration prior knowledge and the quantity of models to approximate the score distribution will also be investigated. Combinations of these shall also be illustrated.

3.3 Randomise

Figure 3.3 show the degree of variability for the 10sec-10sec evaluation with a variety of T-Norm cohorts, albeit engineered with extreme scenarios. To show statistical significance of a random cohort, we shall investigate the consequences by randomly selecting 50 impostors from pool P illustrated in Figure 3.4(a), 3.4(b), 3.4(c) and 3.4(d) respectively for the 10sec-10sec, 10sec-1conv, 1conv-10sec and 1conv-1conv 2005 NIST conditions. Though not a logical approach in an application scenario,



(a) 10sec-10sec

(b) 10sec-1conv

(c) 1conv-10sec

(d) 1conv-1conv

Figure 3.4: An illustration of the effects of 50 cohorts containing 50 impostor models, selected at random for different evaluation conditions in NIST 2005

it is plausible to use a random procedure to select impostors for T-Norm with sub-optimal results.

As expected, variability is a issue with different cohorts in all task conditions, though higher variability is shown with targets trained on short utterances. For the 10sec short utterance enrolment, the maximum deviation between the extremes of the EER is 3.1% for 10sec-1conv testing and 4% for 10sec-10sec trials. Similarly, for both the 1conv-10sec and 1conv-1conv evaluations, the

difference in EER is approximately 0.5% and 0.7%. It has been shown here that the test utterance duration contributes to overall SV performance (i.e. the relative performance of the 10sec-1conv giving a performance range between 24.99% to 28.14% EER is of greater robustness than the 10sec-10sec which provides performance in the range of 32.59% to 36.57%). We can observe that the performance variability from the 10sec trained targets show higher sensitivity towards the contents of the T-Norm impostor cohort, contrast to well trained 1conv targets.

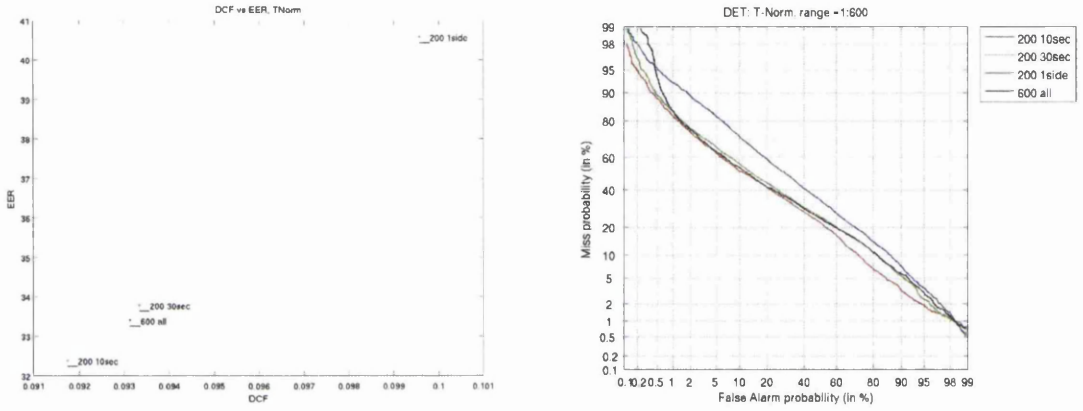
3.4 Matching Cohorts through Prior Knowledge

As discussed earlier in Chapter 1.1, prior knowledge is a tool that we can exploit which may enhance the application of T-Norm by selecting impostors for its normalisation cohort. This section emphasises on applying prior knowledge to compose an impostor cohort for T-Norm.

Since the NIST 2004 evaluations, the idea of matching T-Norm cohorts has been known. Here, this supposition will be investigated by applying matched and miss-matched speakers (defined in introduction of this chapter) to different evaluation scenarios. The makeup of pool P encompasses impostor models from three conditions of training, defined by the amount of target-specific data. For certain evaluations we have possible matched and miss-matched impostors that resemble the target models. For the 10sec and 1conv training conditions we shall investigate the effect of the T-Norm when applying 200 impostors from the three cohort categories, 10sec, 30sec and 1conv. We shall use all 200 same-conditioned impostors as the normalisation cohort to provide a greater Gaussian approximation from the available models for a more stable statistical representation. The results are presented as a performance measure between experiments based on their EER and DCF, accompanied by the tradition DET plots.

For the 10sec-10sec performance depicted by Figure 3.5, the matched cohort provides best results when applying T-Norm, with an EER of 32.39% and DCF of 0.09. For a cohort composition of 200 miss-matched 1conv impostors, a drop of approximately 8% EER and 10% for the DCF is illustrated. From the performance attributes shown on the DCF vs EER plot, the all pool cohort provides close performance to the 10sec cohort for both EER and DCF, though at both extremes of the DET curve 3.5(b) the all pool performance is lacking. From observations of Auckenthaler [13] (and the artificial DET plots shown in the appendix, Figure A.2) it is intuitive to state that a decent performance increase is provided by the all pool cohort comes from the influence of the standard deviation component. If a straight line was plot from the two extremes of the all pool DET plot, the performance would be close to that of the miss-matched 1conv pool, where the mean of both the true and false score distributions become closer, ideally illustrated in Figure A.2(a). The close intersection of the all pool and 10sec measures shown in the plot is a positive contribution from

the change in variance to both true and false score distributions. For high-acceptance applications that reside with low Miss probability and high false probability, the matched impostors provide superior robustness. The scenario to consider is no prior knowledge is available for a specific selection criteria would be an all pool cohort. However, this can encompass longer computation and also, as shown in this short duration scenario, inferior results. If prior knowledge is available, it should be utilised.



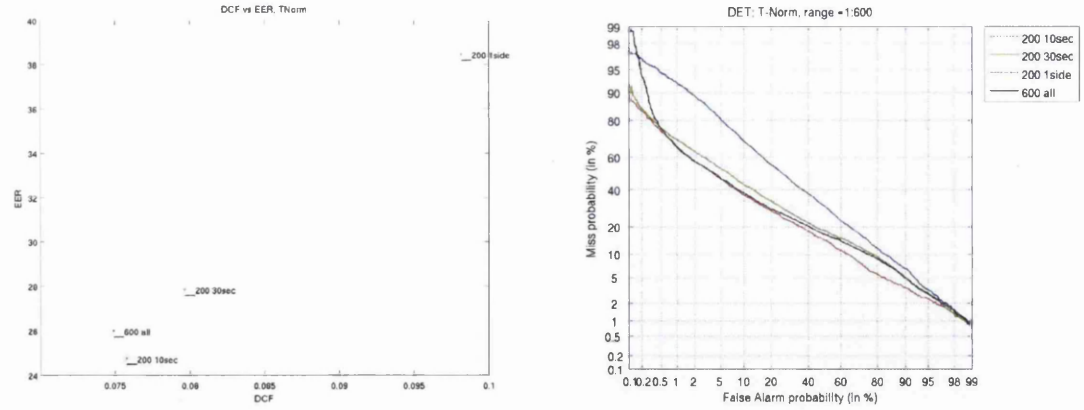
(a) (b)

Figure 3.5: An illustration of matched and miss-matched for the 10sec-10sec NIST 2005 evaluation performance, depicted by DCF vs EER (a) (x-axis and y-axis respectively), with the DET plot (b) for matched and miss-matched cohorts.

Similar performance attributes are illustrated with the 10sec-1conv condition (Figure 3.6 with categorised impostor cohorts by their duration). Again the target-matched 10sec cohort provides best outcome. Applying more test data (2.5 minutes over 10sec) the EER of the matched cohort gives an overall 8% enhancement, though the high miss-matched 1conv cohort only provides a 2% improvement over the short duration test utterances. The *all pool* composition shows similar results to the 10sec cohort with a deviation of 1.2% for the EER from the matched cohort. Yet again, on both extremes the high acceptance (region 3 of the DET curve) and high security (region 1 of the performance curves) the matched cohort prevails. So far, this is true for both 10sec target training conditions.

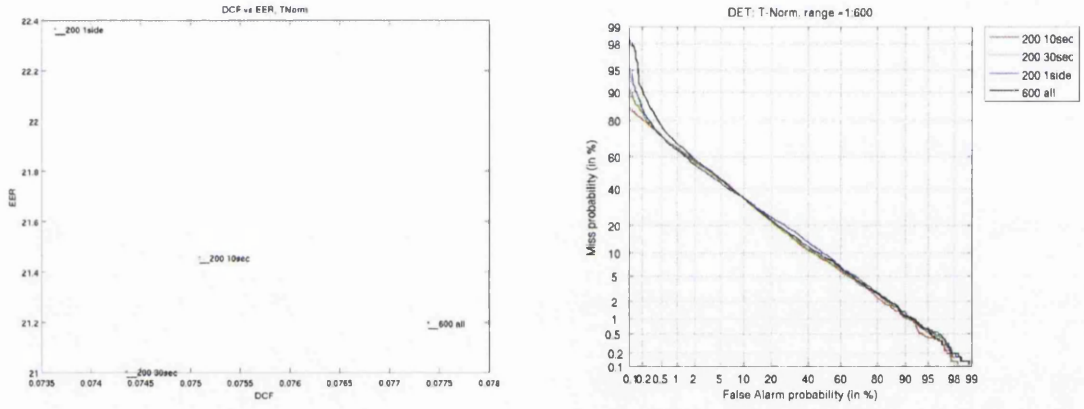
The 1conv training scenarios tell a different story. Increasingly, the greater training data had an unexpected effect. For the 1conv-10sec scenario illustrated by Figure 3.7, all cohort combinations of the all-pool, matched and miss-matched have similar portrayals. Primarily, linear characteristics are shown for each plot with negligible differences between either cohort, though 30sec miss-matched cohort gives approximately 1% advantage in both EER and DCF.

Figure 3.8 displays better system performance with increased utterance duration for both test and train, the 1conv-1conv does not require a specific cohort for meaningful enhancement. Again,



(a) (b)

Figure 3.6: A depiction of scores from the 10sec-1conv NIST 2005 evaluation performance, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b) for matched and miss-matched cohorts.

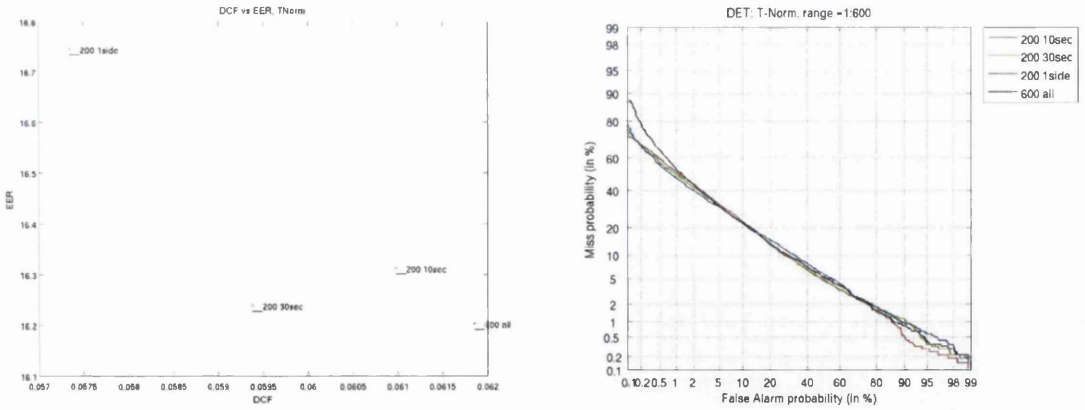


(a) (b)

Figure 3.7: An illustration of scores for matched and miss-matched cohorts for the 1conv-10sec NIST 2005 evaluation performance, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and the DET plot (b).

linear plot characteristics are shown with a tolerance of less than 0.5% EER between them.

From the matched cohorts experiments, we can see that the short-duration tasks require some prior knowledge to be applied for better performance when selecting the normalisation cohort for T-Norm. With a low quality 10sec mode, the sensitivity of T-Norm is shown. Observations catered from the 10sec training conditions show similar degradation to that observed by Dunn et al. [42] when H-Norm statistics were applied to miss-matched training and test utterances. The use of the 1conv cohorts shows high detrimental performance in the 10sec scenario, comparable to the high variability of performance of the low bit rate encoding of the MELP coder displayed in [42]. From



(a) (b)

Figure 3.8: An illustration of scores for the 1conv-1conv NIST 2005 evaluation performance when applying matched and miss-matched normalisation cohorts for T-Norm, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and the DET plot in (b).

the reported variation between evaluation conditions of the 10sec and 1conv, the reported GSM coder in [42] of lower compression displays lesser variation with miss-matched H-Norm cohorts, vaguely analogous to the 1conv training scenario of higher speaker information and relative insensitivity to miss-matched cohorts for T-Norm.

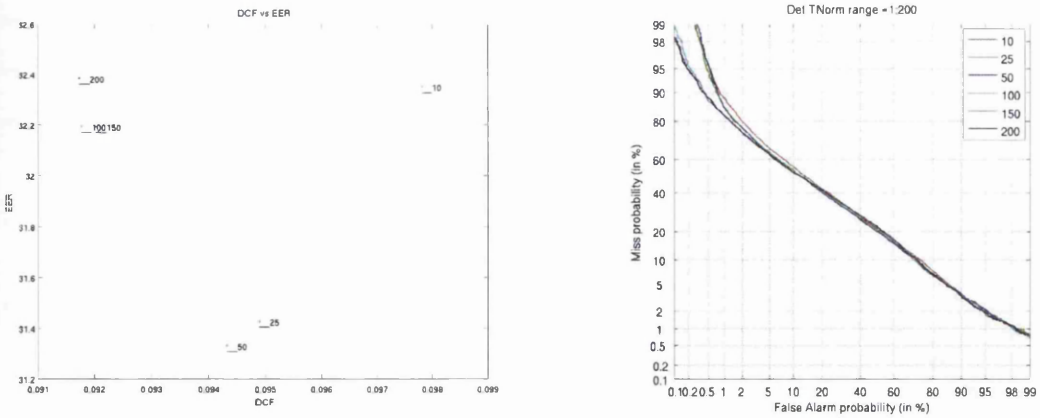
With all of pool P, a variety of matched and miss-matched models provide decent performance at the EER and chosen DCF. It is interesting from these experiments to observe the lack of sensitivity for the long duration task with higher-quality models with respect to the cohort chosen. This suggests that a component within the T-Norm approach degrades the performance with specific models during the short duration task. It has been observed by Fauve et al. [24] that the features designed for utterance durations of 2.5 minutes (1conv) are suboptimal for that of 10sec. Such features, coupled with score normalisation domain ramifications could of course be a path of investigation, though it will not be considered in this work. The configuration of T-Norm has been shown to follow a condition specific requirement.

Following the observations of published impostor quantities and for completeness, we shall examine the effect of modifying the quantity of impostors to populate the cohort.

3.5 Impostor Cohort Quantity

Mixed observations have been published by Roland [10] using a quantity of 50, commented by Navrati and Ramaswamy [11] to use 100 impostor models and Sturim and Reynolds [12] including 55 models to their cohort. Here, we shall attempt to clarify the requirement by investigate the outcome

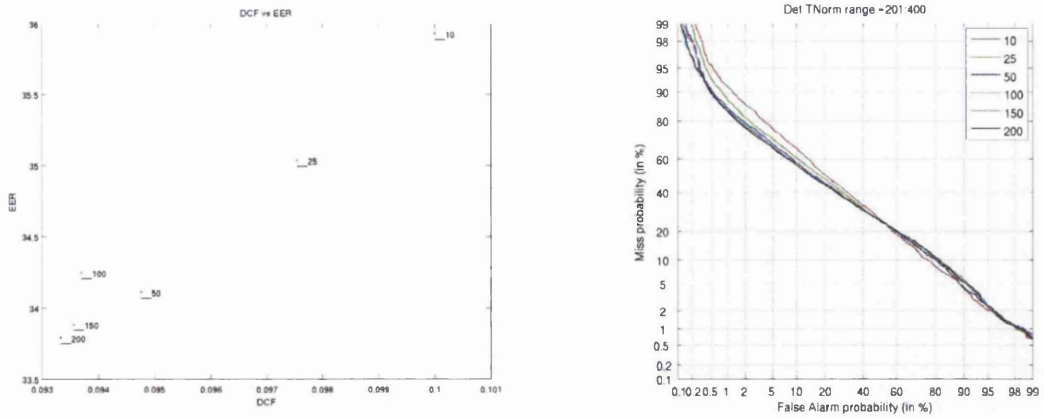
when the quantity of impostors that define a cohort changes over different task durations. The following normalisation cohort experiments will focus on the short-duration task with confirmation in the longer-task evaluations. We shall begin the cohort size investigation by initially concentrating on the 10sec-10sec condition of NIST 2005; also comparing the matched and miss-matched T-Norm sets to determine if the size of the cohort influences performance. It is hypothesised that a once a set of impostor models reaches a certain threshold, further increase of the impostor set would not provide change to the statistical mean or standard deviation. The range of cohort sizes considered are 10, 25, 50, 100, 150 and the full 200 for each of the utterance-duration defined impostor sets.



(a) (b)

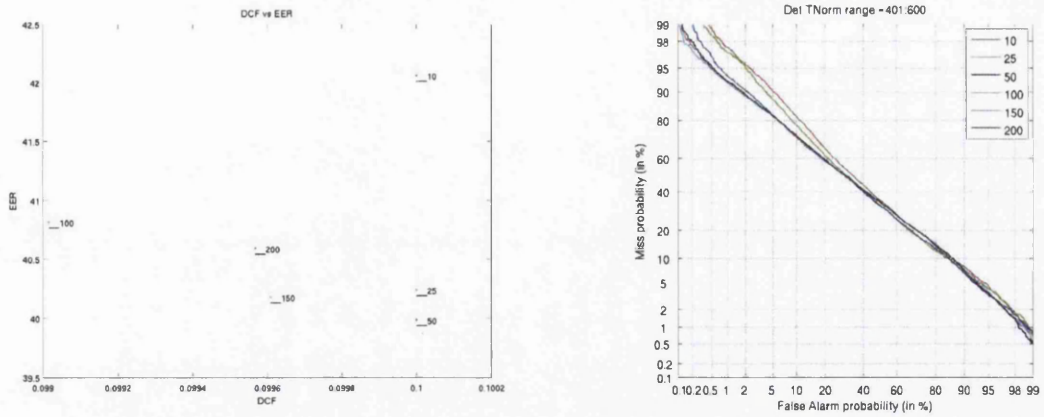
Figure 3.9: An illustration of performance by changing the quantity of models whilst confining to a matched 10sec within the cohort. Subfigures show 10sec-10sec NIST 2005 evaluation performance, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b).

Figure 3.9 depicts a collection of performance plots for a matched cohort T-Norm with the 10sec-10sec evaluation when changing the quantity of impostor models to generate the normalisation statistics. For this condition, the best EER is provided with 50 T-Norms as reported by Auckenthaler and similarly by Sturim and Reynolds, though the DCF of both cohorts of size 25 and 50 give marginally better EER. Once 100 models are used, the change in performance becomes negligible. It is to be noted that here, only one collection per size is considered and as we have shown from the random test, an error factor of approximately 4% is expected when selecting a certain combination of impostors for a given cohort. Of course, some poorly derived impostor models could influence a small selection of impostors.



(a) (b)

Figure 3.10: An illustration of performance by changing the quantity of models within the cohort, whilst miss-matching to a 30sec cohort. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b).



(a) (b)

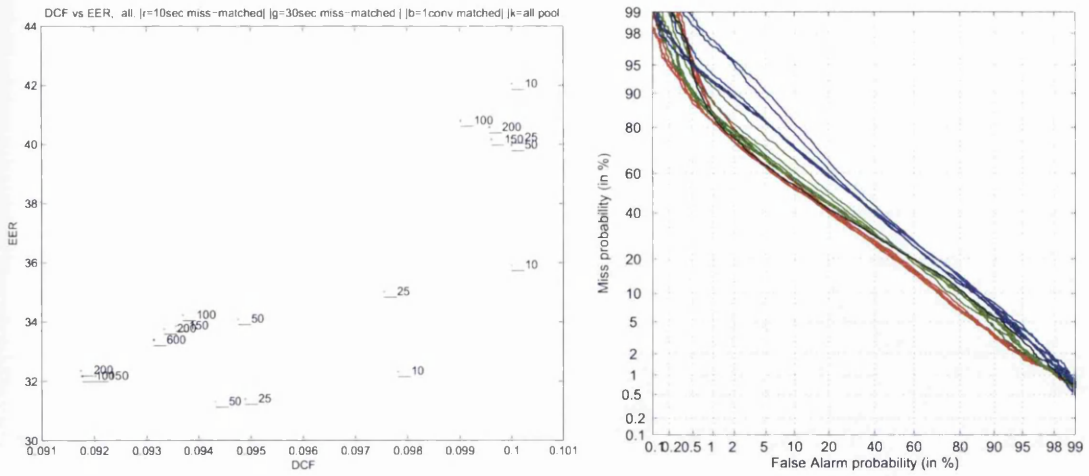
Figure 3.11: An illustration of scores by changing the quantity of models within the cohort, whilst miss-matching to a 1conv cohort. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted by DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b).

For the same evaluation condition but substituting for a 30sec miss-matched cohort, Figure 3.10 a tolerance of approx 1% EER and 1.5% DCF when changing increasing from a population size of 50 to 200, though 200 is marginally better. There is just under 3% EER between the best choice from the 10sec and 30sec cohorts.
For the highly miss-matched T-Norm set, shown in Figure 3.11, the DCF has less than a 1% change between the higher capacity cohorts. The cohorts below 100 show poor performance in the high security domain and maximised the cost function to 100%. There is approximately 9% difference in EER robustness between this miss-matched and the 10sec matched cohort.

Cohort	EER for 10sec		EER for 30sec		EER for 1conv	
quantity (k)	min %	max %	min %	max %	min %	max %
10	31.4	34.7	34.4	37.9	41.3	44.7
25	31.2	33.8	33.5	37.5	40.05	44.3
50	31.2	33.55	33.4	36.5	39.9	44.1
100	31.2	33	33.3	36.2	40.4	44.2
150	31.2	33	33.3	36.3	40.1	43.5
200	31.39	31.39	33.75	33.75	40.6	40.6

Table 3.2: This table illustrates the range of performance provided by the minimum and maximum EER from the 10sec-10sec NIST 2005 data set when 50 evaluations are performed per size/matching combination

Exhibiting only one performance curve for the lower cohort sizes does not provide statistical confidence. Table 3.2 shows the maximum and minimum EER for the 10sec-10sec evaluation with different cohort quantities. In general, this illustrates the more models the less deviation. However, highlighted by the minimum EER for the matched scenario, there are cases when the cohorts of less quantity provide better performance, indicating an impostor model, or sub-set of, providing unreliable observations to the normalising distribution. In the all pool case, selection is limited causing unreliable models to be used to in the normalisation distribution, for models of 10sec training, unreliability of models is greater as little speech is available to a given speaker.

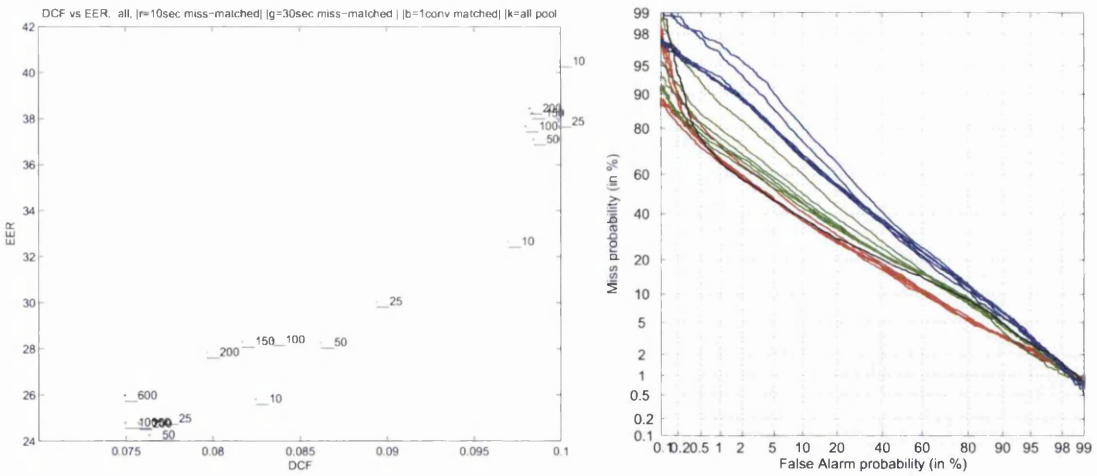


(a) DCF vs EER

(b) DET plot

Figure 3.12: The conventional DCF vs. EER is shown in (a) and DET performance plot in (b). Here a collection of performance plots for a matched and miss-matched T-Norm evaluation of 10sec-10sec are shown with different amounts of impostor models used to generate the represent the normalisation cohort. The red = 10sec matched, green = 30sec miss-matched, blue = 1conv miss-matched and black = all 600 models.

Figure 3.12 demonstrates the combination of previously discussed results into a single illustration. The red plots show the outcome when using a matched T-Norm cohort and the green and blue highlight the miss-matched 30sec and 1conv cohorts respectively. Generally from Figure 3.12(a), we can observe that an increase above 100 models in the T-Norm cohort generally do not effect EER performance. All distributions show a decrease of DCF when the cohort volume is lower or equal to 50. The all pooling 600 T-Norms have moderate performance, slightly better than the miss-matched 30sec condition. There is a trade-off between efficiency and performance regarding the quantity of impostor models used, but for the 10sec trained models, we can deduce that the size of the cohort should generally not be less than 100 for robust performance.

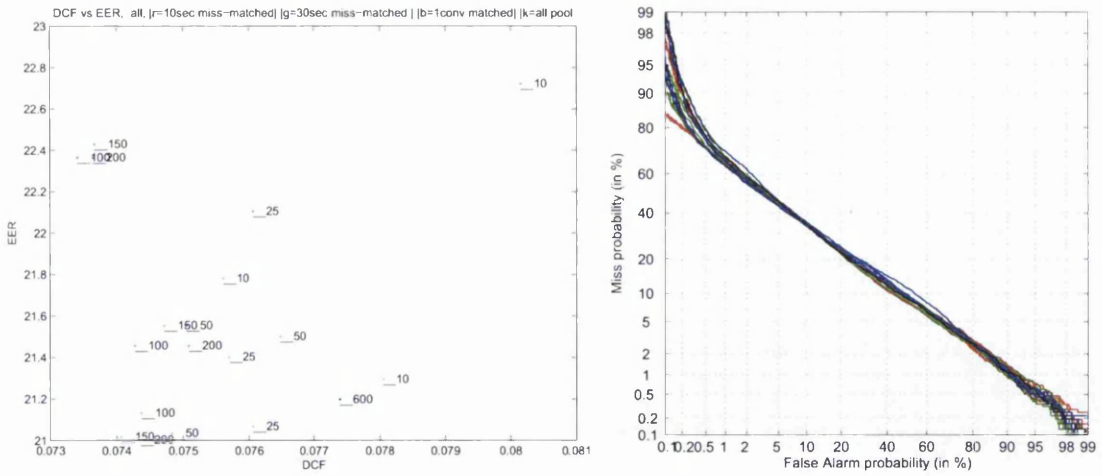


(a) DCF vs EER

(b) DET plot

Figure 3.13: The conventional DCF vs. EER is shown in (a) and DET performance plot in (b). A collection of performance plots for a matched and miss-matched T-Norm evaluation of 10sec-1conv rather than the 10sec-10sec evaluation of Figure 3.12. Different amounts of impostor models used to generate the statistics, red = 10sec matched, green = 30sec miss-matched, blue = 1conv miss-matched and black = all 600 models.

Similarly to the 10sec-10sec condition, the general rule of selecting a sub-set of matched impostors that is greater than or equal to 50 applies. This gives similar performance when applying a mixed set of 600 models (performance dictated by label (_600) on Figure 3.13(a)). On an efficiency basis, a reduction in computation of 91.7% is shown for a matched cohort of 50 impostors over the all pool P. There is a certain sensitivity when selecting impostors for the 10sec training conditions.

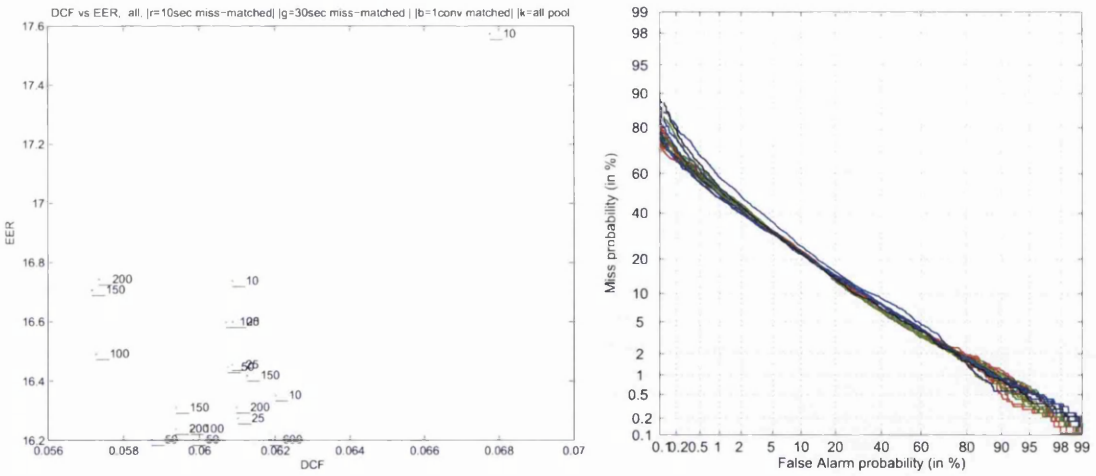


(a) DCF vs EER

(b) DET plot

Figure 3.14: The conventional DCF vs. EER is shown in (a) and DET performance plot in (b). A collection of performance plots for a matched and miss-matched T-Norm evaluation of 1conv-10sec with different amounts of impostor models used to generate the statistics, red = 10sec matched, green = 30sec miss-matched, blue = 1conv miss-matched and black = all 600 models.

However, the necessity of the matching rule is reduced with the 1conv training conditions. By examining the 1conv-10sec results first, depicted by Figure 3.14 as described in the same condition for the previous experiment, the use of either matched or miss-matched conditions do not provide meaningfully better results. A 1% deviation is shown when transferring between matched, miss-matched or the all pool cohort.



(a) DCF vs EER

(b) DET plot

Figure 3.15: The conventional DCF vs. EER is shown in (a) and DET performance plot in (b). A collection of performance plots for a matched and miss-matched T-Norm evaluation of 1conv-1conv rather than the 1conv-10sec evaluation of Figure 3.14. Different amounts of impostor models used to generate the statistics, red = 10sec matched, green = 30sec miss-matched, blue = 1conv miss-matched and black = all 600 models.

For the 1conv-1conv condition, the choice of quantity of impostors from 25 to 200 does not show the same detrimental trend. Whilst applying only 25 models for normalisation can further increase the computational efficiency with negligible performance loss. Again 10 T-Norms provides worst results with any of the matched or miss-matched cohorts. It is interesting to note that the all pooled 600 T-Norms (the black plot) again shows good EER but similarly a poorer DCF from the degradation in the high security region of the DET plot.

3.6 T-Norm Component Analysis

Attributes from the matched/miss-matching experiment for obtaining impostor statistics influence the final verification performance, some combinations hindering while others are equal or better than the baseline UBM-LLR. Naturally, cohort size is an important factor too, providing smoother distribution statistics with a higher gross of impostors. Though, contrary to this observation, the matching/miss-matching criteria is an interesting situation where, as an overall performance, the statistical parameters derived from the all pool scenario shows poorer results over the matched scenario.

Certain impostor models could cause this, though within each of the impostor cohort combinations shown, the statistical significance from one of the contributing components, i.e. mean or standard deviation, may provide a detrimental influence.

With the conventional statistics, we shall investigate the contribution of the different T-Norm components from the distribution scaling equation 2.5. We ask the question, during test normalisation, does either component (i.e. the mean and/or standard deviation) degrade performance? We have conformed to a publication from Auckenthaler et al. [10], a cohort size threshold of greater than or equal to 50 is required for decent performance with the 10sec training condition and here, found that generally 25 or more for the 1conv training condition. We have observed in Figure 3.9 that an odd taper in both the high security and high acceptance domains of the DET plot when using an all pool scenario which should, conceptually lead to a better impostor approximation. We have seen that a cohort quantity below a certain threshold also shows this phenomena, though only in the high security region. Initially we shall investigate the component contribution with a cohort of 200 models for the three defined impostor pools.

We shall first summarise the breakdown of the distribution scaling approach when applied to T-Norm. Recall from Chapter 2, the LLR for both target and impostor contain the UBM biasing influence. For each target LLR score, derived in equation 2.3, there exists a train of impostor LLR scores illustrated in equation 2.11. The mean and a standard deviation statistics for normalisation are derived from the impostor train.[1]

T-Norm (repeated for completeness)

$$S_{T-Norm} = \frac{\Lambda(B)_T - \mu\{\Lambda(B)_{I_{1\ldots N}}\}}{\sigma\{\Lambda(B)_{I_{1\ldots N}}\}} \tag{3.1}$$

Zero mean

$$S_{T-Normonlymean} = \frac{\Lambda(B)_T - 0}{\sigma\{\Lambda(B)_{I_{1\ldots N}}\}} \tag{3.2}$$

Unity standard deviation

$$S_{T-NormSD} = \frac{\Lambda(B)_T - \mu\{\Lambda(B)_{I_{1\ldots N}}\}}{1} \tag{3.3}$$

To investigate the contribution of the statistical contributions from the distribution equation over various conditions, we shall reduce both the influence of the mean and standard deviation independently. First, this is achieved by setting the mean to an artificial zero for all trials (equation 3.3). The second scenario consists of unity standard deviation across all trials (equation 3.2). Conventional T-Norm and UBM normalisation will also be illustrated. For the following experiments, beginning with 10sec-10sec, we shall derive the UBM score as the baseline (no specific impostor influence). Conventional T-Norm is also used.

[1]This could be replaced with a train of utterances from the same target providing the availability of such speech.

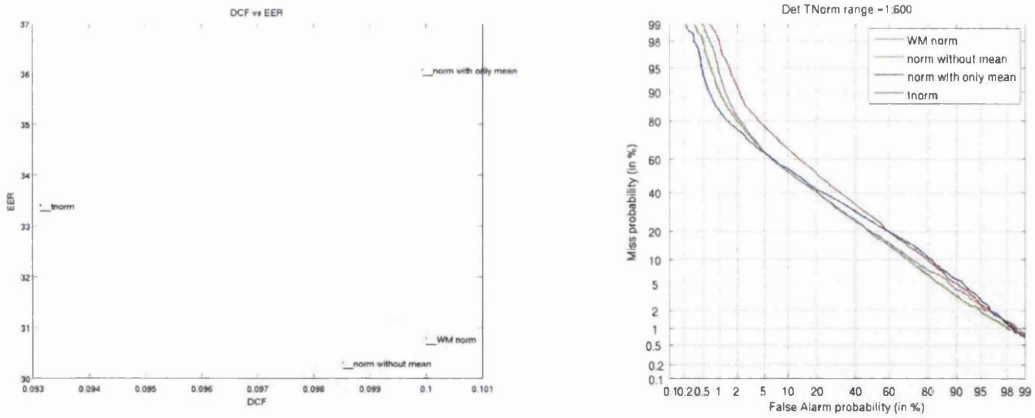
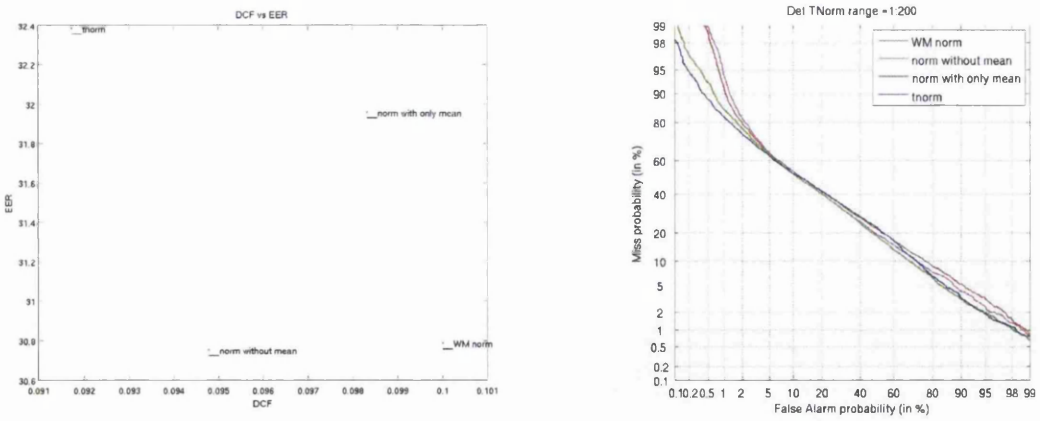


Figure 3.16: Influence of different T-Norm components using pool P for 10sec-10sec NIST 2005, Red = T-Norm with unity standard deviation, Green = T-Norm without mean component, Blue = conventional T-Norm and magenta = UBM LLR. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted as DCF vs EER (a)(x-axis and y-axis respectively) and the conventional DET plot (b).

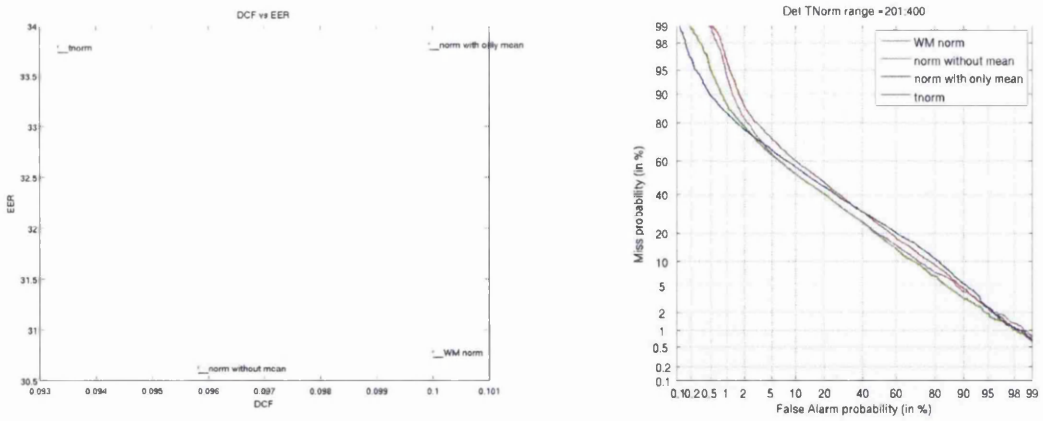
Initially, all of pool P represents the cohort. This demonstrates the breakdown of the T-Norm from its conventional form with traditional world model approach performances, illustrated in Figure 3.16. Here, the conventional T-Norm approach is represented as the blue curve, showing the 'tilt', observed by Auckenthaler et al. [10] against the world model LLR. It is interesting to see the that the T-Norm with only the standard deviation component is similar and slightly better than the world of speech components as this T-Norm composition embodies the UBM LLR within the target score; in addition to the standard deviation. A additional 0.5% is supplied to the EER performance. There is greater negative impact on the mean-only T-Norm composition with an approximate 5% EER loss in the derivation with only the standard deviation.



(a) (b)

Figure 3.17: Influence of different T-Norm components for 10sec-10sec NIST 2005 using only matched share of pool P, Red = T-Norm with unity standard deviation, Green = T-Norm without mean component, Blue = conventional T-Norm and magenta = UBM LLR. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted as DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b).

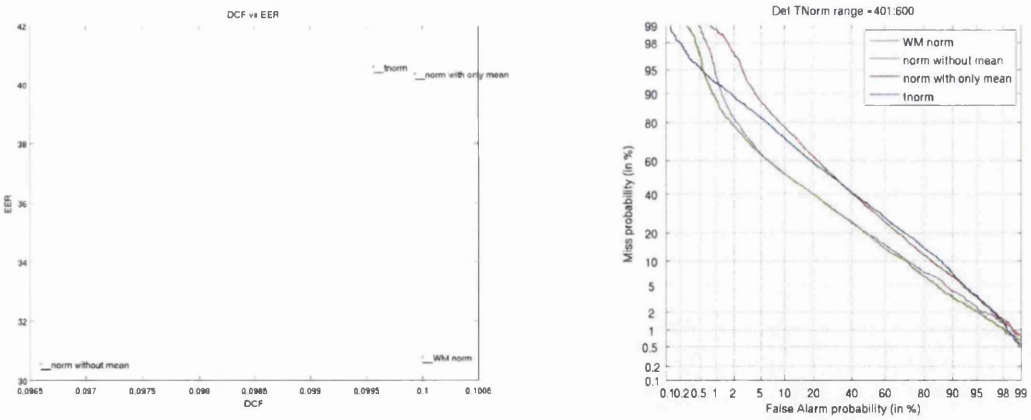
As previously shown, it was beneficial to attempt a matched impostor-to-target situation to obtain greater robustness. It is logical to carry out this observation to the T-Norm component analysis, where Figure 3.17 shows that the overall performance is slightly enhanced in the EER domain with the standard deviation only approach, though this time following the trend of the conventional T-Norm (highlighted in Figure 3.17(b)). Again, the mean component appears to hinder the T-Norm with just over 1% EER loss to the standard deviation only condition.



(a) (b)

Figure 3.18: Scores depicting the influence of different T-Norm components for 10sec-10sec NIST 2005 using only miss-matched 30sec share of pool P, Red = T-Norm with unity standard deviation, Green = T-Norm without mean component, Blue = conventional T-Norm and magenta = UBM LLR. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted as DCF vs EER (a)(x-axis and y-axis respectively) and the conventional DET plot (b).

When analysing the miss-matched scenarios, a reduction of performance was expected as previously shown during matching and miss-matching experiments. The baseline UBM score has the same characteristic does not change from previous experiments (no impostor cohort influence). It is surprising to show that the standard deviation scenario provides virtually the same results as the 10sec matched and all pool scenario for the EER attribute. There is a variation of approximately 0.04 DCF. It is intriguing to observe an almost static performance level using a miss-matched set with the target LLR and only the standard deviation statistic of the impostors. Finally we shall see if this attribute carries forward to a highly miss-matched cohort of 1conv impostors. It is hypothesised that the large variation of 1conv impostor scores, as seen in Figure 2.2 with a performance shown in Figure 3.18 would cause larger statistical miss-match.



(a) (b)

Figure 3.19: The influence of different T-Norm components for 10sec-10sec NIST 2005 using only miss-matched 1conv share of pool P, Red = T-Norm with unity standard deviation, Green = T-Norm without mean component, Blue = conventional T-Norm and magenta = UBM LLR. Figures show 10sec-10sec NIST 2005 evaluation performance, depicted as DCF vs EER (a)(x-axis and y-axis respectively) and DET plot (b).

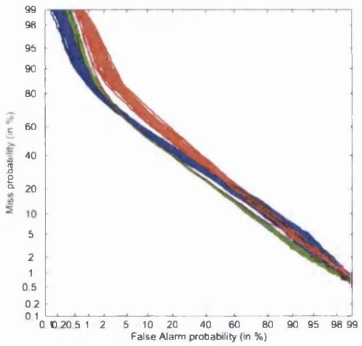
As previously illustrated, the 1conv miss-matched scenario causes performance degradation with conventional T-Norm, with the mean-only T-Norm following the poor performance trend. However, the standard deviation only T-Norm again surprisingly enhances on the baseline of the UBM (magenta on Figure 3.19). This unexpected result highlights a problem with the distribution scaling approach used for T-Norm when not taking care with T-Norm selection. There is around a 10% EER loss when conducting conventional T-Norm with highly miss-matched impostors, illustrating that the mean as the culprit of performance degradation at the EER with respect to the short-duration task.

For future investigation, the use of other statistical measures, such as median or mode could provide more robust influence when performing such impostor-centric score normalisation.

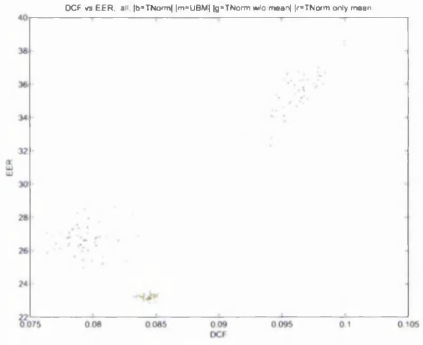
3.7 Component Analysis with Random Selection

To provide confidence of the results with the different T-Norm parameters, the previous results will be repeated with 50 random selections of impostors; similar to that in section 3.3.

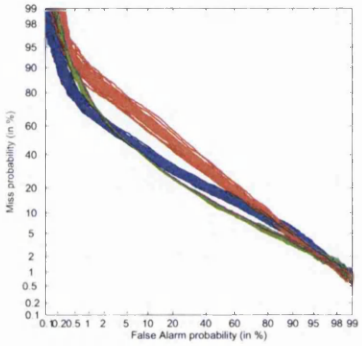
(a) 10sec-10sec



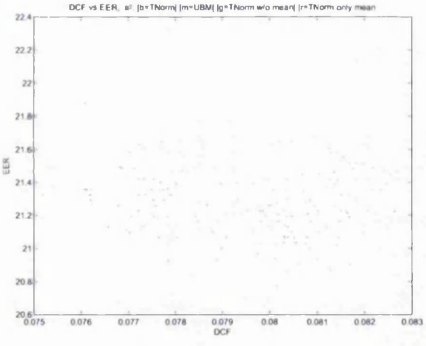
(b) 10sec-10sec



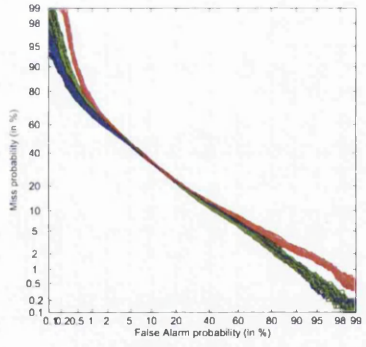
(c) 10sec-1conv



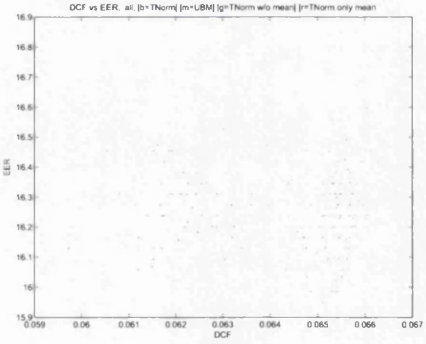
(d) 10sec-1conv



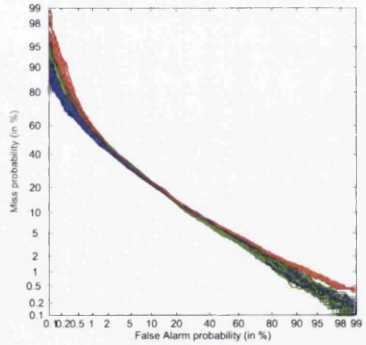
(e) 1conv-10sec



(f) 1conv-10sec



(g) 1conv-1conv



(h) 1conv-1conv

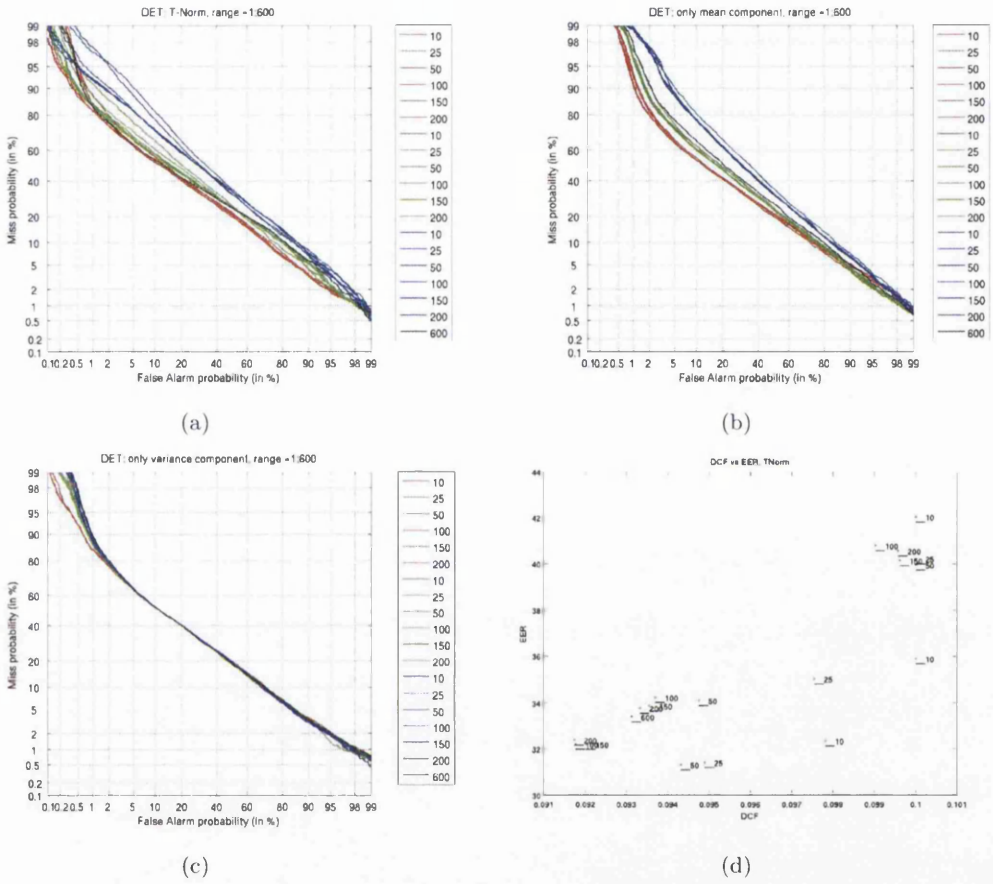
Figure 3.20: Scores of 50 plots per illustration with randomly chosen impostors under different task conditions represented as a DCF vs. EER plot and DET plot where the blue illustrates conventional T-Norm configuration, the single magenta in each graph represents UBM normalisation; green shows the T-Norm with only the standard deviation component and the red plots shows the T-Norm with only the mean component.

The red profiles show the peformance of the impostor-centric scaling with only the mean component. The green plots depict the performance with the world normalised score, scaled by the standard deviation. The blue profile illustrates the conventional T-Norm approach, a fusion of the mean and standard deviation components. The 10sec training conditions show that the variety of scores are when different T-Norm equation configurations are applied showing the higher sensitivity of the individual T-Norm normalisation components when short training data is used for enrolment. This is less variable when different test utterance durations are applied. This sensitivity is reduced when target models are trained with greater amounts of speech, depicted by the 1conv training conditions in Figures 3.20(e)- 3.20(h). The deviation of the mean for the short-duration 10sec trained models display the poor performance result as explained in section 3.6. Again, the short-duration models with cohorts of alternate contents, the T-Norm with only the standard deviation approach shows consistency at the EER with a variety of cohort compositions. It is observed that the general nature of the standard deviation only approach shows a less scattered performance nature in shorter enrolments, Figures 3.20(a) to 3.20(d). This approach appears to be more robust to the composition of the impostor normalisation cohort than the mean only configuration. It appears that the mean statistic has greater variation in both the EER and DCF, especially in the short duration evaluations. However, the mean only component is depicted in 3.20(e) to 3.20(h) provides similar EER performance though degraded DCF with longer target enrolment.

From these results, it appears that the mean component is highly sensitive and is a degrative contribution to the T-Norm. This is emphasised greater in the shorter enrolment evaluations.

3.8 Component Analysis with Different Cohort Size

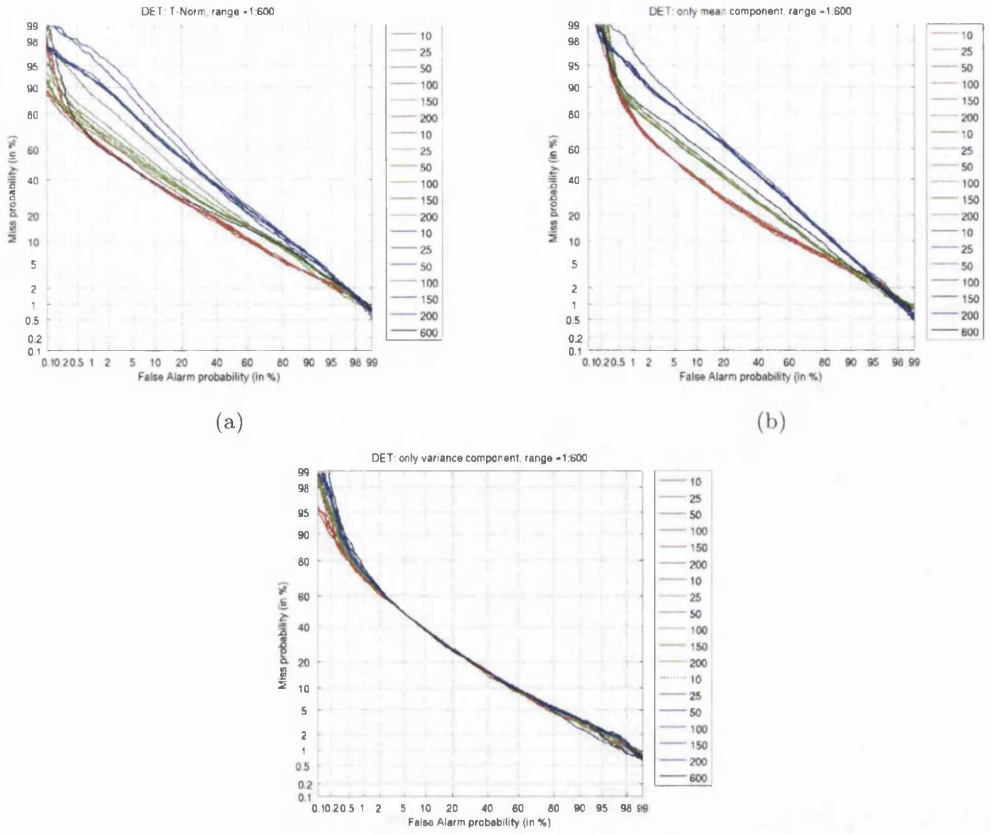
Figure 3.21 a, b and c displays the different T-Norm configurations with the equivalent DCF vs EER plot shown by Figure 3.21(d). It is intuitive to view the degrading performance of the mean-only T-Norm configuration when greater miss-match is applied (figure 3.21(b)). There is little change when applying the standard-deviation only concept to any matching combination, illustrated in Figure 3.21(c). The miss-matched 1conv (blue) degraded performance shown on Figure 3.21(a), must be a result of the degraded effect of the impostor mean contribution.



(a) (b)

(c) (d)

Figure 3.21: Short task evaluation illustrating, (a) the variability of T-Norm when changing both size (from 10 to 200) and match condition. Figure (b) highlights the variability caused by the mean only component and (c) demonstrating the small deviation when only using the standard deviation only component. The composition of the DCF vs. EER performance is shown in (d). The colours are representative of the cohort composition from pool P where red, green and blue represent impostors trained on 10sec, 30sec and 1conv durations. The black profile shows conventional UBM normalisation performance.



(a) (b)

(c)

Figure 3.22: 10sec-1conv task with colours being representative of a selected cohort composition from pool P, Figure 3.2. Figure (a) depicts the variability of T-Norm when changing both size (from 10 to 200) and match condition. (b) mean only component and (c) the standard deviation only component. The composition of the DCF vs. EER performance is shown in (d). Again, red, green and blue represent impostors trained on 10sec, 30sec and 1conv durations. The black profile shows conventional UBM normalisation performance.

The 10sec-1conv NIST 2005 evaluation provides further uncertainty to the viability of the mean for short-task durations. A divergence of approx. 17% EER is displayed by the mean component with different matched conditions; influenced by the conventional T-Norm with an approximate 14% EER deviation. Again the standard deviation only scenario has less than a 1% EER change. The best overall performance is supplied by the matched 10sec impostors with conventional T-Norm; a difference of 1.23% EER in favour of the standard deviation only approach and 0.01 DCF enhancement with the conventional T-Norm approach. Here, the conventional T-Norm appears to provide a linear DET curve, indicating that the true and false distributions are following a Gaussian trend. The 1conv training evaluations for both 10sec and 1conv show a different story; results shown in Figures 3.23 and 3.24 respectively. The 1conv-10sec demonstrates negligible performance (less than 1% EER and 0.01 DCF) between the different combinations of T-Norm and the conventional UBM-LLR. It is clearer to see the straightening of the DET curve is a contribution of the standard

deviation scaling contribution. A 0.05% difference resides between the best EER performance of 21% provided by the conventional T-Norm and the equivalent standard deviation only setup.

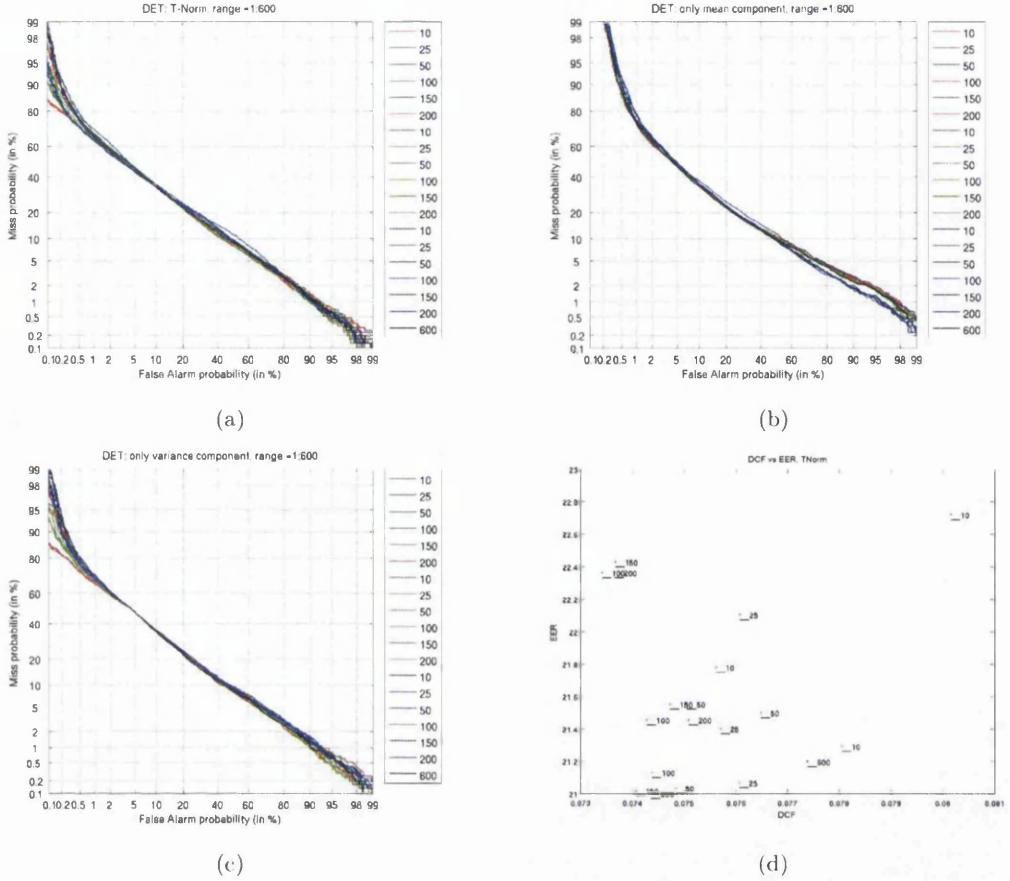


Figure 3.23: 1conv-10sec evaluation illustrating, (a) the variability of T-Norm when changing both size (from 10 to 200) and match condition. (b) mean only component and (c) the standard deviation only component. The composition of the DCF vs. EER performance is shown in (d). Colours are representative of the cohort composition from pool P, Figure 3.2, red, green and blue represent impostors from 10sec, 30sec and 1conv durations. The black profile shows conventional UBM normalisation performance.

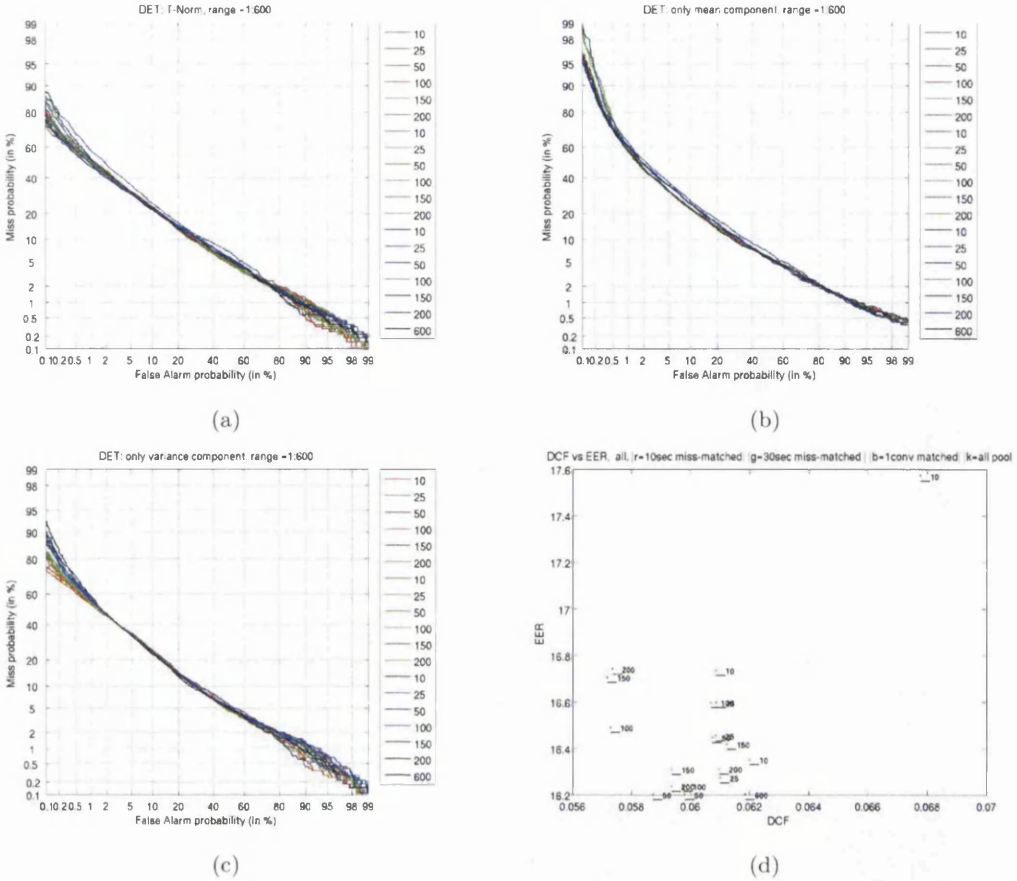


Figure 3.24: 1conv-1conv evaluation illustrating, (a) the variability of T-Norm when changing both size (from 10 to 200) and match condition. (b) mean only component and (c) the standard deviation only component. The composition of the DCF vs. EER performance is shown in (d). Colours are representative of the cohort composition from pool P, Figure 3.2, where red, green and blue represent impostors trained on 10sec, 30sec and 1conv durations. The black profile shows conventional UBM normalisation performance.

3.9 Discussion

In this chapter we have discussed the sensitive nature of the T-Norm approach. Arbitrarily choosing different impostor models as a general T-Norm set for all trials, it observed that the EER can range around 13%, highlighting the variability of selection. Limiting the T-Norm selection using prior knowledge, with formulation of a matched target to impostor approach gives better performance. In the short-duration task, there is a high degree of sensitivity for the 10sec models target in relation to the mean component of the T-Norm. Where as the scaling characteristic of the standard deviation for short-duration trained evaluations has greater benefit for EER peformance. From results examined in this chapter, it appears that the mean component is highly sensitive and is a degrative contribution to the T-Norm, especially in evaluations with short training utterances.

Substituting the T-Norm approach with a new statistic in place of the mean may be beneficial for the low trained target models. However, it is not so for the 1conv case, by showing greater robustness with a diversity of cohorts when using the T-Norm approach. The all pool configuration, consisting of a mixture of 600 models give admirable results in general over all task durations. However, in applications with limited resources, using such a large cohort is difficult to consider.

Poor performance of the conventional T-Norm approach for the 10sec training task may be caused by suboptimal features to represent the utterances. Fauve et al. [24] shows a suboptimal performance through the SVM approach when the same features are applied on different evaluation conditions. Also, the current MAP adaptation approach with a lack of training data may contribute to the unreliability of the impostor models.

It has been shown in the literature [43] that these techniques perform better for a similarity domain when the cohort is selected in a trial-dependent way. The next chapter will investigate this by conducting T-Norm with trial-dependent cohorts. Approaches to provide a target-dependent cohort of impostor models have been published for T-Norm by Sturim & Reynolds [12] coined, adaptive T-Norm (AT-Norm) and also a Kullback-Leibler measure by Ramos-Castro et al. [40] which will be discussed in the following chapter. It would appear that the benefit of matching a subset of impostor through the T-Norm approach, the aphorism "there's no data like more data" only holds true in finite cohorts with non-detrimental impostors.

All the work conducted here could be extended to a bilateral scoring approach with likely enhanced performance.

Chapter 4

Trial-dependent Selection for T-Norm

As discussed in section 3.5, performance of the 10sec-10sec evaluation with a matched impostor cohort of 200 speakers (also highlighted in Table 3.2) can show a surprisingly degradative performance over a smaller sub-set of impostors to represent a cohort. This suggests an obtrusion by certain impostor models on the normalising statistics. Removing the influence of such hindering models provides more robust normalisation statistics. This chapter furthers the discussions of T-Norm cohort selection approaches from the previous chapter by selecting a cohort of impostors independently, for each target model.

4.1 Trial-dependent T-Norm

Finan el al. [32] discuss impostor cohort selection approaches in the text-dependent speaker verification (SV) domain using models constructed with VQ templates. Using a distance metric between a true and false score development distribution to investigate the size of normalisation cohort for all target models in subsequent trials. As shown in the previous chapter, one approach is to select models that best represent the targets modelled on a certain duration described by the evaluation condition. An automatic option comes from an enhancement to traditional T-Norm, introduced by Sturim & Reynolds in 2005 [12]. This is known as the *adaptive T-Norm* (AT-Norm) approach to set a more robust decision threshold and also assists with computational performance in the scoring stage. Here, each target is tied to a sub-set of impostor models, selected through a derived criteria; subsequently used for T-Norm with trials assessed with the particular target model. Consequently providing a further robustness with scoring by discarding impostors that hinder the statistical make-up of the normalising cohort. Selective score normalisation approaches such as, *adaptive T-Norm* and *Kullback-Leibler T-norm* (KL-TNorm) [40] are used to provide an automated target-specific selection process. These approaches are shown to give enhanced results over the fixed trial-independent (or target/speaker-independent) cohort of impostors applied to all target models of an evaluation.

Again the same impostor-centric normalisation method is applied, modifying the individual hypothesised scores to a common trial scale, in an attempt to reduce effects of environmental mismatch between target model and test utterance.

4.2 Adaptive T-Norm (AT-Norm)

AT-Norm, is a data-driven approach, utilising an additional pool of speech utterances as passive data (or data drivers[1]) to generate a score distribution of false trials. The resultant distribution is assumed to only contain false outcomes, as the data-drivers (extra data for a data-driven approach) are assumed to originate from the pool of *other* prior knowledge, described in Chapter 2, where the utterances are different to that of the target and impostor models. Again, the size of the potential T-Norm impostor cohort is E. The pool of data-drivers are of N quantity. The target-specific model is scored against all of pool N, deriving a collection of scores, perceived as a 1-dimensional vector of size N. The same pool of passive data is then scored against each impostor model in the T-Norm cohort, again resulting in a 1-dimensional vector of size N for each impostor model. In total, we have one target vector of N coefficients and E impostor vectors of the same order, depicted in Figure 4.1.

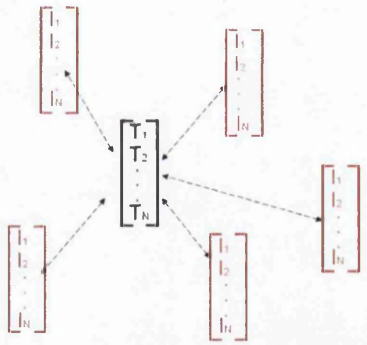


Figure 4.1: Adaptive T-Norm selection process with E impostors, illustrated as red vectors derived by N scores, represented by the number of vector coefficients, here depicted as five red N dimensional vectors.

To select impostors, a vector distance metric is used to observe the proximity of each impostor to the target. Sturim & Reynolds apply the $L-Norm$ to select (equation 4.1).

$$D(T - I_e)_{L-Norm} = \sum_{n=1}^{N} |(T_n - I_n)| \tag{4.1}$$

[1]The utterances provide stimuli to the impostor and target models to drive out a score

To gather a relevant subset of $\bar{E}$ imposter models for a given target, the impostors are ranked by a nearest neighbour criteria. The closest $\bar{E}$ impostor models are then tied to the particular target model. This selection method is repeated for each enrolled model during training. This approach is deemed adaptive as a subset $\bar{E}$ from the impostor pool has been tailored to the particular target model. During all subsequent real trials, the test-normalisation mean and standard deviation statistics for a given target model are only generated by the subset of impostors, chosen through the adaptive selection procedure.
The experimental trend reported by Sturim & Reynolds in [12] depicts a general performance increase when selection has occurred with an impostor cohort, usually containing models that is considered to match the target under trial. Again, matched to their training utterance duration (NIST training condition). They also show that a pool of impostors containing a selection of training conditions (NIST 1conv, 3conv and 8conv) show slightly worse performance than a matched impostor cohort. From experiments conducted here in the development and evaluation trials showed this is not true, as some models, especially in the 1conv task tended to include some 30sec based models and forcing a matching criteria deteriorated results by approximately 1% EER.

One disadvantage of an adaptive approach for selecting speaker-specific impostors is the application of the in-situ adaptation of speaker models. When a speaker model has been verified as a true speaker for a given trial, the model can be enhanced by further adaptation using the test utterance as a continuum of training data. When the model has been further adapted, re-selection would logically be required. Though for the evaluation-dependent cohorts, this was found not to significantly degrade the performance with well-trained targets and 10sec impostor normalisation cohort. This is true for both AT-Norm and the following approach.

4.3 KL T-Norm

Ramos-Castro et al. [40] describes another approach to selecting target-dependent T-Norm. Here, a distortion measure is generated between target and impostor models using the Kullback-Leibler (KL) divergence criterion. Similarly to AT-Norm, the KL approach apples a proximity ranking methodology to select a subset of appropriate T-Norm impostors. The KL-TNorm takes a non-empirical approach to select the cohort by only utilising a divergence criteria between the target and each impostor. A model to model comparison by measuring the distance between the Gaussian components of a target model to their equivalent Gaussian component in each impostor. The KL divergence observes the distortion of two probability density functions.

$$D\left(f\middle|\hat{f}\right) \equiv \int \left(\frac{f}{\hat{f}}\right) \tag{4.2}$$

The correspondence of $f_i = \hat{f}_j$ is assumed when $i = j$ as both GMM distributions come from the same UBM and no MAP adaptation has taken place on the i^{th} component. Furthermore,

diagonal co-variances are used per mixture component. The computation of the divergence metric for speaker models can be represented by equation 4.3.

$$\int\left(\frac{f}{\hat{f}}\right) = \frac{1}{2}\left[(\mu_i - \hat{\mu}_i)\hat{\Sigma}_i^{-1}(\mu_i - \hat{\mu}_i)\right] \tag{4.3}$$

The advantage of this method is that no extra data is required to stimulate the impostor cohort and target model for selection. Experimental approaches to the KL-TNorm approach proved unsuccessful and from the published results in [40], the KL selection approach over their baseline results show less relative enhancement than the AT-Norm [12] approach. From this, we shall therefore conduct a further analysis of T-Norm on the *adaptive* selection approach.

4.4 Experiments with the Adaptive T-Norm (AT-Norm)

Published results by Sturim and Reynolds [12] show positive enhancement to SV when selecting impostors for evaluations with longer-duration. Although, it was reported that negligible change is shown in the short-duration task, 10sec-10sec, it is postulated that no results have been presented when the selection pool includes models trained on low amounts of speech to provide a performance gain in these conditions. This will be examined in this chapter along with the standard 1conv conditions for a comparison of relative enhancement. The same impostor pool P applies with the passive data-drivers obtained from the test utterances used in the NIST 2004 evaluations.

Impostor selection experiments based on the procedure described in section 4.2 is presented with the use of an example 10sec target model from the NIST 2005 10sec-10sec evaluation. The blue plot on Figure 4.2 shows the accumulated distortion measure of E impostor models to a single 10sec trained target model when scored against N passive-utterances, where E is 600 and N is approximetly 500. The x-axis represents the impostor model identity (representative of the order is described by Figure 3.2) and the y-axis represents the accumulated scores of the N data-drivers. There is a clear three-step trend with a majority of the 10sec impostors giving closest proximity, whilst 30sec domain (model IDs 201 to 400) miss-matched impostors have less similarity and more so with 1conv impostors. The three steps are representative of the three domains of residing impostors from pool P 10sec, 30sec and 1conv trained models. All but one of the 1conv impostors (models indexed from 401 to 600) demonstrate high similarity to the 10sec trained target. This *outlier* was found to be a poorly trained 1conv model, where most feature vectors samples have been filtered by the speech detector. The red plot indicates the ranked proximity to the 10sec target model, again the three 'steps' can be seen. Figure 4.3 now extends to show the impostor proximity over all 10sec targets for the 10sec-10sec evaluation in NIST 2005. The colours are representative of models trained with an approximation of utterance duration, representative of pool P (described in Chapter 3). The x-axis represents the impostor model ranking from 1 to 600 where index 1 is the

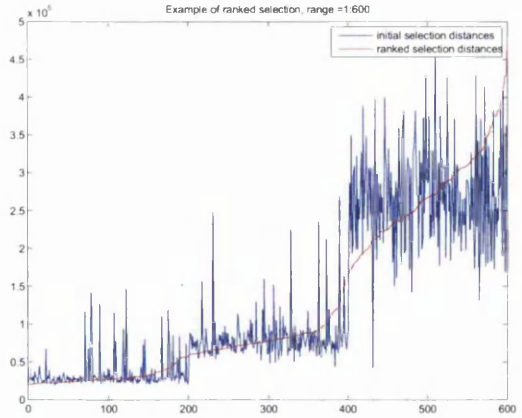


Figure 4.2: The x-axis shows 600 impostor models, in the 10sec, 30sec and 1conv ID configurations rated through distortion measures from a single target model represented on the y-axis in the 10sec-10sec evaluation, the red plot indicates the model distances in a ranked order. This ranked order is shown for over 250 targets in Figure 4.3.

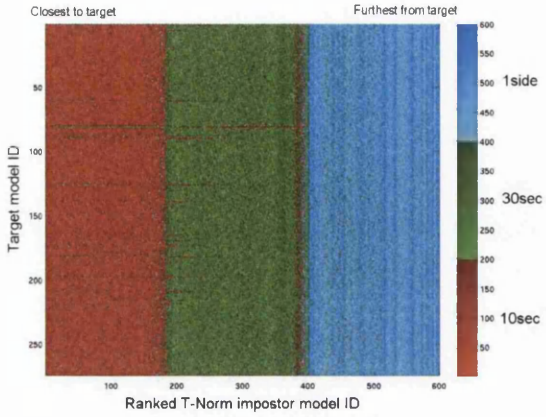
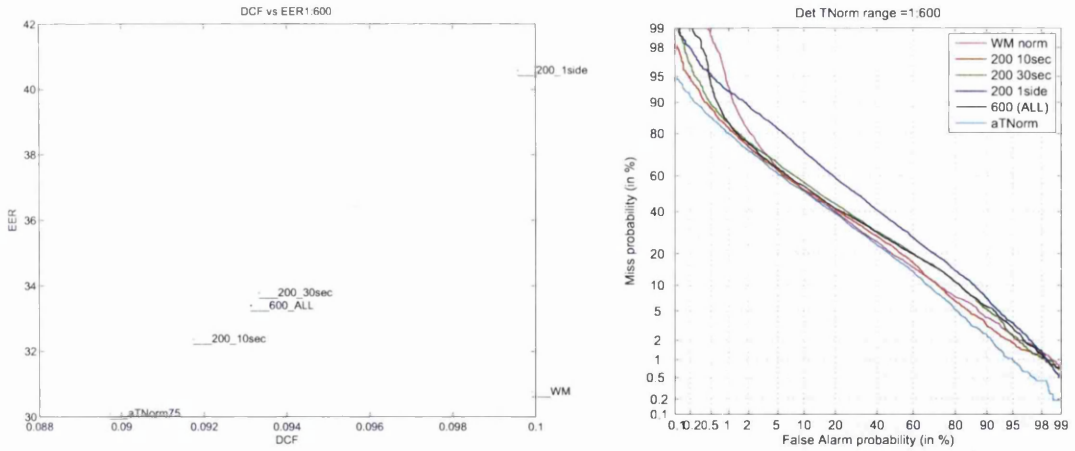


Figure 4.3: The impostor ranking for each target model used in the 10sec-10sec evaluation is shown by an arbirary ID on the y-axis. The colour represents the impostor model origin from within pool *P*. Here, the x-axis represents the ranked cohort order of the impostor models that have been adapted towards each target. Impostor models with the left most rank represents close resemblance (better matching) to the target, whilst the rightmost impostors are deemed miss-matched.

closest to the target model. The target model is identified by the y-axis, in the order supplied by the NIST training index. The ranked impostors depicted in Figure 4.2 by the red plot, represents a single row of ranked impostors in Figure 4.3. For AT-Norm, we can select the number of impostors to use from 1 to all 600. For example, if the AT-Norm criteria was to select the 50 closest impostors to the target ID's; impostors ranked 1 through to 50 would be linked to the appropriate target model. For all 10sec targets, 98.56% of the impostor models originate from the 10sec portion of pool *P*. Selecting all 600 impostors would give the same result when using the all pooled T-Norm cohort as shown in the previous chapter. Here, approximately 10% of the 30sec impostor models are collected within the first 200 indexed models (models between 180 and 200), indicating that around 10% of the 10sec models are considered more of a miss-match to evaluations containing targets of 10sec training.

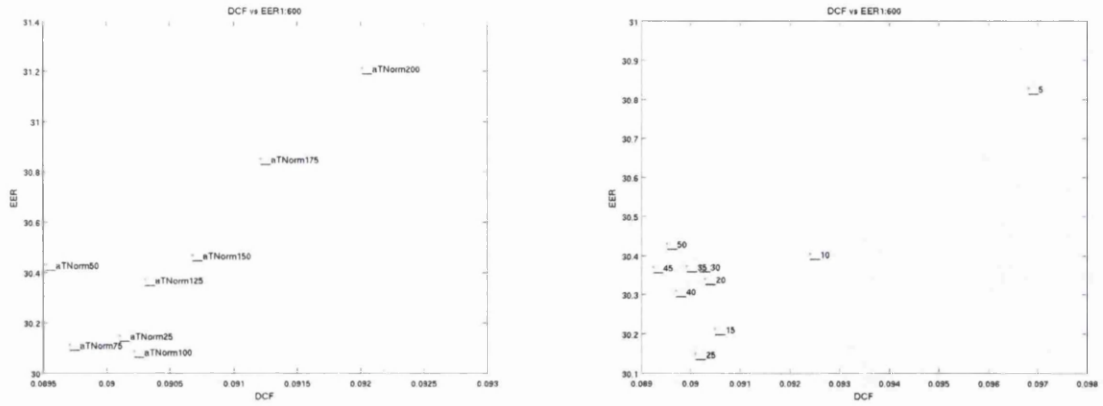
It was found that applying the initial 75 of the ranked impostor for T-Norm gave increased performance over the matching approach in the evaluation-specific scenario; the AT-Norm ranked impostors approach for the NIST 2005 10sec-10sec is shown in Figure 4.4. A decrease in EER and DCF of approximately 2% is shown over the matched, evaluation-specific 10sec cohort.



(a) (b)

Figure 4.4: 10sec-10sec NIST 2005 performance with AT-Norm with 75 impostors, presented by DCF vs. EER in (a) and the DET performance (b)

Figure 4.4(b) shows that the adaptive approach towards model selection gives a greater linearity of the performance curve. The adaptive approach has proportionally tightened the variance of both true and false distributions. Specifically providing a matching in the AT-Norm is no longer required as this approach established a matched scenario automatically for the majority of targets.

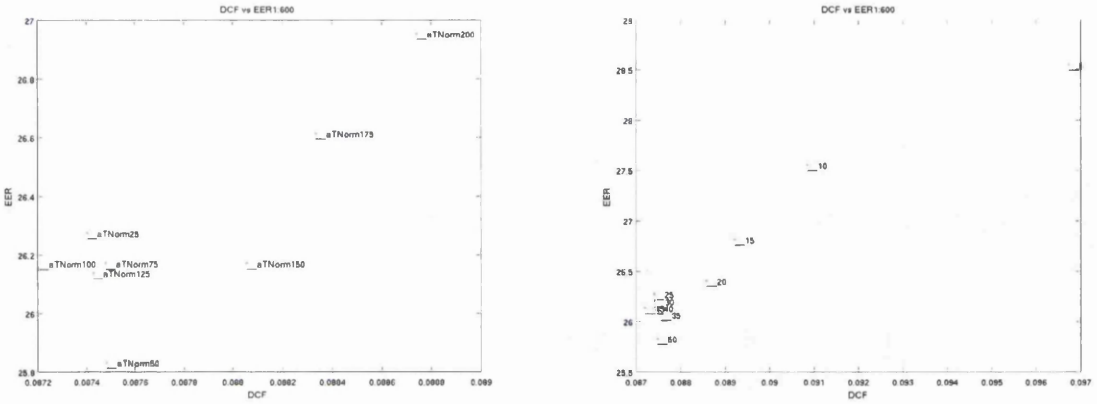


(a) Range 25-200

(b) Range 5-50

Figure 4.5: 10sec-10sec NIST 2005 performance with AT-Norm, presented by EER vs DCF with a coarse range between 25 and 200 in (a) and a finer resolution with low impostor quantities in (b)

Figure 4.5 shows the outcome of the AT-Norm approach when the selection parameter $\bar{E}$ is varied coarsely from 25 to 200 in 25 increments and in a finer resolution of between 5 and 50 in steps of 5. The result og 200 models, shown in Figure 4.5(a) gives improved performance of approximately 0.9% for the EER by replacing the aforementioned *poor* 10% 10sec impostors with 30sec impostors. Decreasing the size of the cohort generally gives a trend of increasing performance. Supplying 75 impostors to each target model provided best overall performance, followed closely by a cohort of 25. It is interesting to observe the scattered nature of the AT-Norm in the short duration task. Supplying a cohort size of between 15 and 50 (shown in Figure 4.5(b)) has a minimal deviation of 0.3% in the EER and a small 1.5% variation in the DCF. A cohort of just 15 models gives a surprisingly admirable result, which would certainly be beneficial in real-world applications.



(a) Range 25-200

(b) Range 5-50

Figure 4.6: 10sec-10sec NIST 2006 performance with AT-Norm, presented by EER vs DCF with a coarse range between 25 and 200 in (a) and a finer resolution with low impostor quantities in (b)

To show confidence in the results, Figure 4.6 presents experimental observations on the NIST 2006 evaluation set. Linear characteristics are evident when the classifier when AT-Norm has been applied, indicating equalised variance for both true and false distributions. Again, low compositions of adaptively selected impostor cohorts show more robust results decreasing from a quantity of 200. The straightened DET curve reported in the NIST 2005 evaluation is also evident here.

From these results, it is shown that the adaptive method on a 10sec trained evaluation assists the performance of SV with cohorts as low as 15. Now, the same procedure will be applied to the longer task duration evaluation of 1conv-1conv. Similarly to Figure 4.2, Figure 4.3 shows the proximity of the impostor models to a single target, trained with a 1conv utterance. Here, we can see that the 10sec impostor domain is being primarily rendered the furthest from the target model, complying with the matching ideology of the evaluation-specific cohort selection. Of course, in an application scenario, impostor searching with the adaptive algorithm could be reduced by initially constricting the selection pool, in this case disregarding the 10sec impostor pool. However, one target model depicted in Figure 4.8 bears close resemblance to a short-utterance trained model by selecting a majority of the personal cohort from the 10sec domain. Here it was found that although a large amount of speech was provided, only 3055 feature vectors were extracted, where a 1conv utterance should consist of around 13000 feature vectors. It was found to be a recording of the expected 5 minutes duration though the content of speech was approximately 20 seconds. It seems that the caller hung up after a total of 34 seconds of telephone conversation. As previously mentioned, Sturim & Reynolds showed that applying a matching criteria gives better selection results quoting

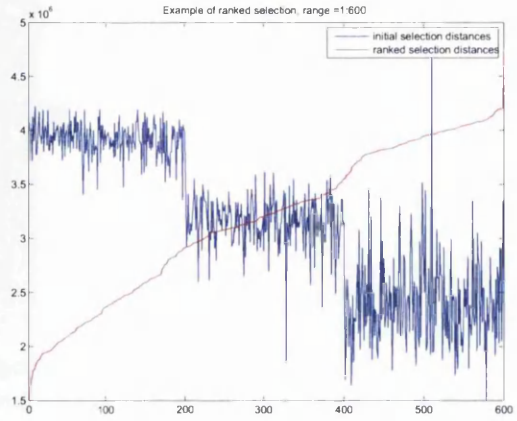


Figure 4.7: 600 impostor model distortion measures from a single target model in the 1conv-1conv evaluation, the red plot indicates the model distances in a ranked order

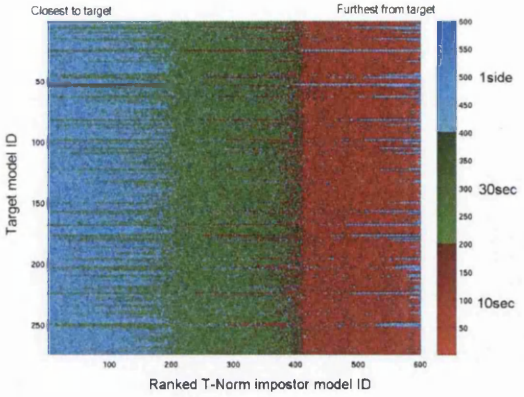


Figure 4.8: The impostor ranking for each target model (y-axis) in the 1conv-1conv evaluation. The colour represents the impostor model origin from within pool P

"Ideally, the pool of cohort models P should be large enough to provide a representative coverage of T-Norm models from which to draw". Though as we can see from Figure 4.3, many 1conv targets utilise the 30sec models and fewer 10sec impostors. Contradicting Stirum & Reynolds' matched observations due to the nature of a subset of target models. This can be caused by unexpected lack or increase of the enrolment speech duration of a speaker and is discussed further in Chapter 5. For a true evaluation, it would appear that a broader set of models trained with a variety of utterance durations would be beneficial to *surround* the assumed training utterance duration. This would account for 1conv models with lower or greater than the count of feature vectors that would normally constitute a 1conv model. Sturim & Reynolds showed that the AT-Norm selection approach did not select pure matched impostors, unless forced through matched pools.

For both 10sec and 1conv models, the notion of a *matched* impostor criteria has been carried over from the evaluation-specific to the speaker-specific, adaptive selection approach. Though some models are observed to reside in the thought to be miss-matched domain, indicating a reduction in the amount of utterance duration assumed to train the target models. These models can be considered as being under-trained in the 1conv task. The term *model quality* could be applied to question the validity of these targets. Again, the three-step trend still illustrates a form of demarcation of the imposter by their utterance duration. With the longer-duration task, shown in Figure 4.9, the AT-Norm clearly outperforms the conventional evaluation-specific T-Norm. Linearity is enhanced and an increased anti-clockwise tilt towards lower Miss probabilities (observed by Auckenthaeler et al. [10]) is shown. This suggests a further constriction of the false score distribution (artificially described in appendix A.3). Lower selective cohort quantities again provide enhanced performance.

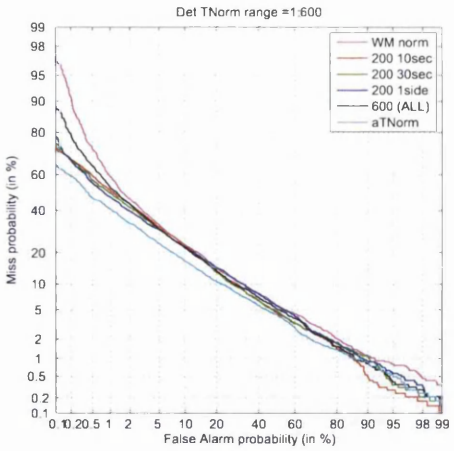
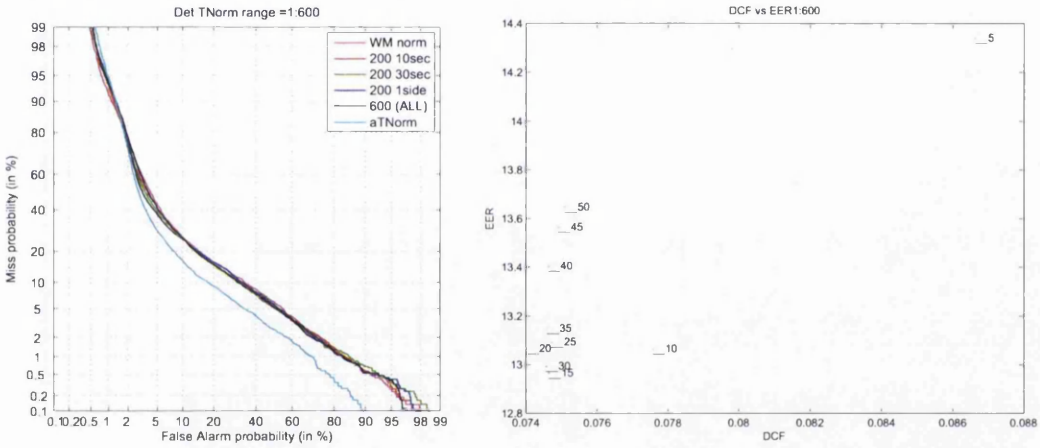


Figure 4.9: AT-Norm (cyan) applied on the 1conv-1conv evaluation from NIST 2005 with illustrations of miss-matched, matched, all pool and basic UBM normalisation.

When the AT-Norm cohort size increases above 50 using impostor models, the contribution of the normalisation statistics decreases. For the NIST 2006 evaluations shown in Figure 4.10, an increase



(a) AT-Norm with 15 impostors per target

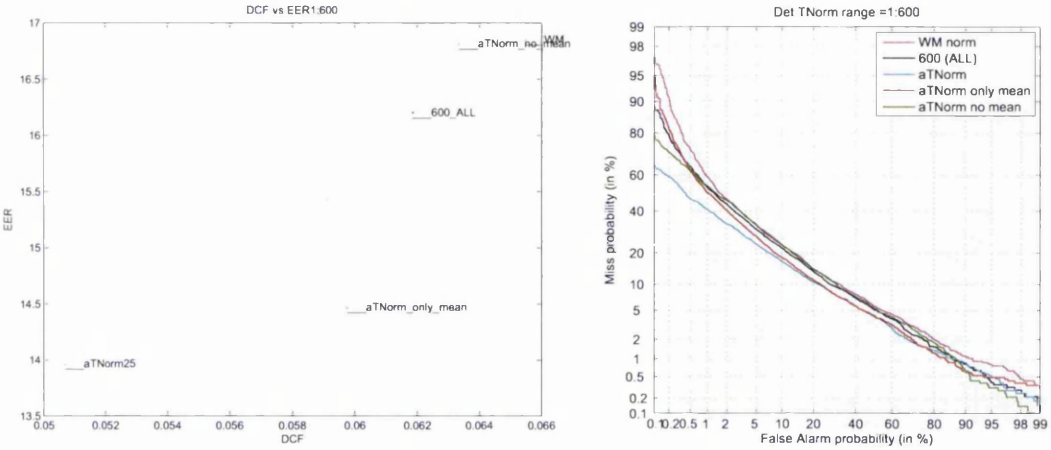
(b) DCF vs. EER with cohort quantity ranging between 5-50

Figure 4.10: 1conv-1conv NIST 2006 performance with AT-Norm (cyan), presented by DET performance with a coarse optimum of just 15 models (a) and the EER vs DCF with different ranges of impostors provided to each target in (b)

can again be seen with the EER for the 1conv-1conv scenario. In contrast to the NIST 2005 evaluations, it was found that a lower cohort size of just 15 impostor models gave the best balanced results for EER and DCF. This could of course be an influence of this particular evaluation database.

4.5 Component Contribution with Adaptive T-Norm (AT-Norm)

In Chapter 3.6, the 10sec-10sec evaluations highlighted the high performance contribution provided by the standard deviation component of the T-Norm. Here we shall determine if this contribution holds true once the models have been tailored to each target. First, the 10sec-10sec task, showing the AT-Norm performance with a selection of $\bar{E}$ 25 with component substitution. Figure 4.11

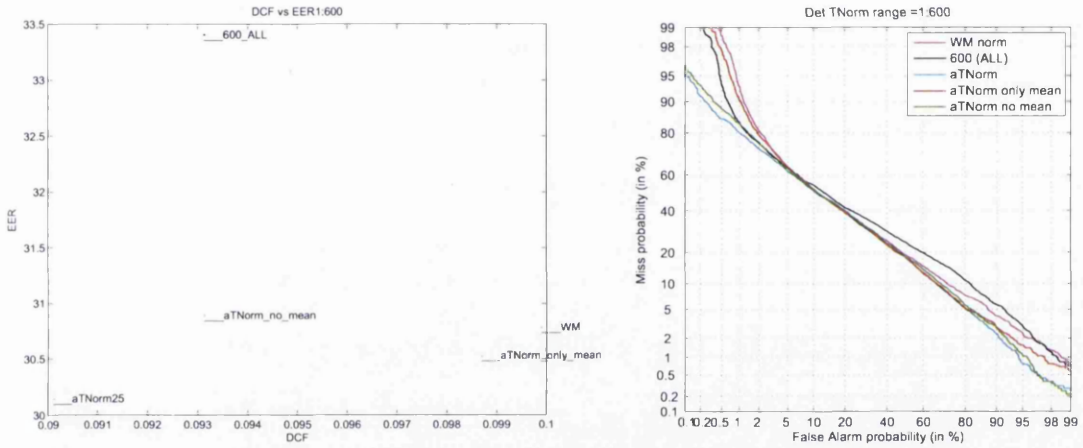


(a) Range 5-50 impostors per target

(b) 25 impostors for AT-Norm (cyan)

Figure 4.11: 1conv-1conv NIST 2005 performance with AT-Norm and distribution scaling component substitution, presented by EER vs. DCF in (a) with different T-Norm components and the DET performance for 25 impostors in (b)

shows the 1conv-1conv AT-Norm performance of 25 impostor models with component substitution. For both the 10sec and 1conv AT-Norm scenarios, the AT-Norm approach now outperforms all combinations of the impostor-centric approach.



(a) Range 5-50

(b) 25 impostors for AT-Norm

Figure 4.12: 10sec-10sec NIST 2005 performance with AT-Norm and distribution scaling component substitution, presented by EER vs. DCF (a) and the DET performance in (b)

4.6 Discussion

In this chapter, approaches to tailor score normalisation statistics through impostor selection has been presented. By supplying each target with its own cohort of impostors using the AT-Norm selection process, a general enhancement to both long and short duration tasks with a minimum contribution of 2% in both EER and DCF was demonstrated. Concurring with results reported by Sturim & Reynolds, the AT-Norm approach enhances the setting of a more robust decision threshold. From results presented in this chapter, this method also applies enhanced performance of approximately 2% to the shorter-duration task over the baseline T-Norm, this is believed to be previously unpublished. The reported use of matched pools for cohort selection with the AT-Norm is not required and as demonstrated by the 1conv trained target experiments; not all target models for a specific condition select impostors that originate from the assumed matched pool. For the short-duration task, lower sized cohorts give best performance gain and secondly, would greatly reduce with trial stage computation. However, more computation is required at target enrolment to select the appropriate impostor models, though this can be processed in an off-line manner. The selection pool needs to be of adequate size and as shown, of training duration variety which to extract impostor models. This was highlighted by Sturim and Reynolds in [12]

For future investigation, the possibility of further enhancement could be achieved by providing a specific cohort size $\bar{E}$ to each target, $\bar{E}_T$. Each target may provide more robust model statistics, not only when the described adaptation approach is applied, but also a customised quantity of impostor models may be assigned to each target. In an unrelated target-centric application, an AT-Norm selection pool could be evolved by tailoring speaker-specific models during the lifetime of a system, by possibly replacing or growing the impostor cohort pool.

An interesting investigation would be to investigate the performance of the aforementioned, Z-Norm, TZ-Norm or the ZT-Norm procedure whilst utilising the trial-dependent 'adaptive' approach discussed in this chapter.

As demonstrated, the tailoring of an impostor cohort to a given target model gives better SV performance. Though, in both the 10sec and 1conv tasks, the nature of a few target models showed false characteristics of originating from other task durations. This was demonstrated by target models that select a majority of impostors considered as being of a miss-matched nature. This question of the target *model quality* will be discussed in the following chapter based from observations of the accumulated impostor model scores generated in the AT-Norm selection process.

Chapter 5

From Model Normalisation to Model Quality

T-Norm impostor models assessed against different test utterances provides a range of scores, as illustrated earlier in Figure 2.3. The cohort can be seen to generally surround the target score over different trial conditions, especially with 10sec target models. Here, the test utterances are considered impostor to both the target and the T-Norm cohort (termed, unfortunately as impostor models). For clarification, in this image we have three sources of information, the target model, the impostor test utterances (stimuli) and the normalisation impostor model cohort used during T-Norm. Under closer examination of these scores, a few impostor models from the normalisation cohort over a set of trials was found to give an unexpected average of high scores. Thus, these few models are hypothesised to always give high scores; in a range of what would be expected from a true target trial, not an impostor trial. Conversely, if a target model displays a high degree of false alarms/acceptances, its reliability, or in this context, its *security* should be questioned before use in subsequent evaluations. In this chapter the measure of a speaker security to a model is discussed, labeled as the *speaker security measure* (SSM), which can be used as a pre-filter, similar to a gender detector, to gain additional knowledge of a given speaker through confidence of individuality against other speaker.

5.1 Introduction

A speaker population can be characterized to contain '*sheep* and *goats*'. Where the *sheep* are usually well behaved and dominate the population, where the *goats* refer to speakers that incur abnormally high amounts of errors and determine overall system performance [3]. Here, the goal is to *'sort the sheep from the goats'*.

The use of quality measures can assist classification by attempting to obtain additional information

about the trial at hand [9], applied with an appropriate action to enhance robustness. Quality analysis can be applied to different stages of the classifier. For example, an initial 'goodness criteria' at the raw signal level through a speech to background noise measure, i.e. SNR. There are a variety of automated assessment procedures to generate such measures. Many approaches are published in the literature to evaluate end-to-end voice quality. *Perceptual evaluation of speech quality* (PESQ) is once such approach which integrates knowledge of the human auditory perception into the scoring procedure. This measure considered state-of-the-art [44]. Here we investigate a targets reliability, postulated from observations from model selection approaches for score normalisation, discussed earlier in this thesis. This gives an insight into the reliability of a speaker, analogous to the alternative hypothesis (H_1) as described in Chapter 2. Koolwaaij et al. [45] concludes that the operation in a real-world application would require an indication of the quality of a newly trained speaker, allowing the system to request re-enrolment or insist on a further contribution of target-specific speech. Motivation from observations on the derived scores through data-driven cohort selection approaches has led to the following work. It is believed that the examination of false scores, generated by assessments of target models against impostors utterance through intentional false trials may yield a measure, descriptive of a target model's quality. If a system can predict the target models that give high errors, the goats, the enrolment procedure can be modified dynamically in an attempt to reduce the errors generated during subsequent trials. One possible procedure is the requirement of additional target-specific speech, though in evaluations such as NIST and some application scenarios, e.g. forensics, this may not be possible.

The quality assessment of target models is not a new area of research. Recently, Richiardi & Drygajlo [9] discuss measures derived from a speech segmentation process in the time domain. A quality measure per vector is derived and can be subsequently used as a weighting factor during target enrolment. Quality measures could also be derived on a feature level, where each feature is assigned a confidence weighting, briefly noted in [46]. A quality measure could also be derived from additional secondary prior information about the training utterances. Richiardi & Drygajlo [9] defines a quality measure as *"a measurable indicator of a factor impacting the classifier behaviour, which exhibits a dependency relationship with the classifier output scores and/or classifier decisions."* Such measures have also been applied at the scoring stage through some score fusion function of the test and/or train speech utterance together with the LLR could also be achieved in the decision process [46, 47]. Weighing factors based on the generated model and signal quality criteria could be used to modify the resultant score. The general system described in [46] generates quality measures to bias the scores for both the model and a the raw test utterance. Koolwaaij et al. [45] obtains indirect measurements by observing the mean and standard deviation from the contribution of both the true utterance scores and impostor based scores obtained from the target model. A weighting regime is applied to subsequent trials. In the method proposed here, we only consider the impostor derived scores. Thompson [5] examines the reliability of a speaker prior

to the model adaptation by examination at the cepstra level. Intra-speaker variation measured through error analysis over short periods between recording sessions in an attempt to classify a speaker as either a sheep or a goat, consequently rejecting or applying a weighting function in later trials.

Model quality assessment is usually applied at the enrolment stage. Other measures can be calculated from the raw signal or at feature level. Following in the theme of this thesis, these types of measures can be considered as a *secondary* form of prior information, an extra component of information that describes the training utterance or model.

As previously discussed, a problem in certain applications is the lack target-specific speech, leading to the use of impostor-centric statistics for score normalisation. A logical approach would be to derive an error analysis of a target through data-driven means, similarly reported by Koolwaaij et al, using utterances assumed to originate from the same target speaker. With the high-security application scenario undertaken by NIST, the availability of extra target specific material is unavailable.

Here, it is hypothesised that any model trialled against an assumed impostor utterance should generate a false observation. These false target scores should fall below a predefined decision threshold during an evaluation resulting in a 'true alarm'/rejection outcome. For a collection of assumed impostor utterances, the statistical mean of an observed target should also give a low average score. If not, it could be deemed a target model of poor quality as many impostor utterances have, on average shown similarity to many other speakers and is likely to generate a high proportion of errors in the false acceptance domain in later trials. The measure described in this thesis could be used to algorithmically reject target models that are likely to cause these errors within the classifier.

5.2 Target quality by false alarm analysis

In general for a data-driven approach there are two categories of data that can be used to determine the confidence of a speaker model, utterances originating from impostors or the target under consideration. Respectively able to derive a confidence of a target's security against impostor attempts or a target's acceptance of utterances from the same speaker. Impostor-centric T-Norm attempts to reduce inter-speaker variations to both target and test utterances by utilising speech that is not of the target, normalising on the false score distribution. This approach is usually taken due to the lack of target-specific and the abundance of impostor speech. The model quality approach described here is applied with the same limited target speech constraint. However, in

some applications where target utterances are common, the latter approach may be accomplished. Similarly to the selection approach during AT-Norm, the application of such data scored against a speaker model must derive some target model characteristic; in this case, a speaker's relationship to impostors.

Here, a collection of many target to impostor scores is used to derive a targets statistical relevance to other impostors. A collection of such scores is presented as a train of scores, represented in equation 5.1. Where E is the number of I impostor utterances and T represents the target model. Here, each score $P(T|I_n)$ in equation 5.1 is normalised by the UBM. Further investigation could apply other score normalisation approaches such as T-Norm. No target reference score exists without deriving a true speaker score from utterances assumed to be spoken by the target speaker.

Here, the *quality* is deemed measurable from the average 'resistance' of a target model to a subset of E impostor utterances drawn from a pool of utterances. In essence this measure is similar to the mean derived for the Z-Norm statistics. Also, SSM observations can be extracted practically during the AT-Norm selection procedure, as the same approach is undertaken to derive score observations for cohort selection. The term *speaker security measure* (SSM) denotes a measurement of a target models resistive nature to impostor attempts, calculated through equation 5.2. The negative sign relates to a natural concept of good and poor quality, respectively with high and low scores. For the duration of this chapter, the potential of the SSM approach is discussed.

$$\{P(T|I_1), P(T|I_2) \ldots, P(T|I_E)\} \tag{5.1}$$

$$SSM(T) = -\frac{\sum_{n=1}^{N} P(T|I_e)}{E} \tag{5.2}$$

A high-level diagram of the SSM is shown in Figure 5.1. Where a model of speaker A is assessed through the SSM approach in parallel with the conventional classification. A weighting regime can then be applied to the classifier outcome with the derived quality measure. If an SSM observation is deemed *low* against a certain criteria, this model could give many false acceptances in subsequent trials. This may assist the setting of a lower, robust threshold. This approach can be deemed as a pre-filter to the verification trial, similar to gender or handset labelling approaches for utterances. This approach aims to provide extra knowledge to assist with making a decision for verification. On a side note, the prior knowledge such as gender is information provided with the NIST evaluations and have not been deduced within the classifier used throughout this thesis.

A *high* score could deem a target to be more robust against false acceptance errors. Preliminary experiments conducted on the SSM will now be discussed.
With all target models, the same data-pool of impostor utterances can be used for assessment to define a distribution of a target population. This is demonstrated here with the 10sec and 1conv

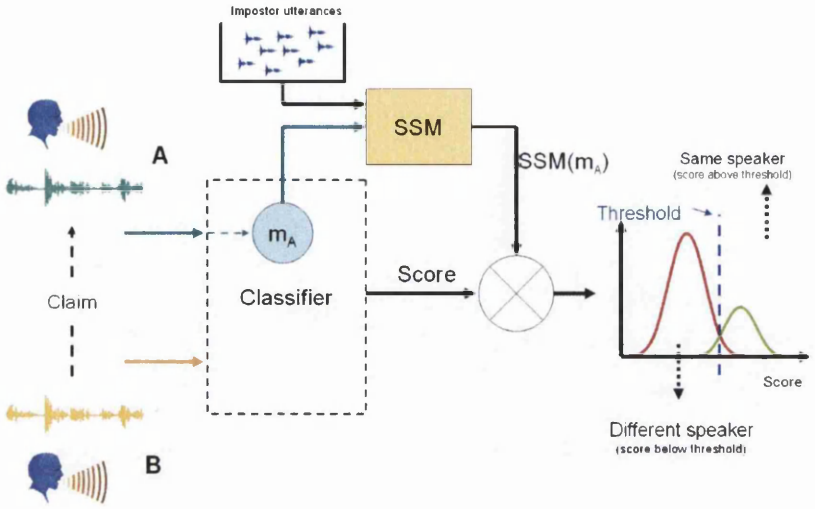


Figure 5.1: High-level speaker classifier view, generating a score. This is weighted with the data-driven SSM approach applied to each target model.

targets used in the NIST 2005 evaluations. In Figure 5.2, the SSM is demonstrated against 274

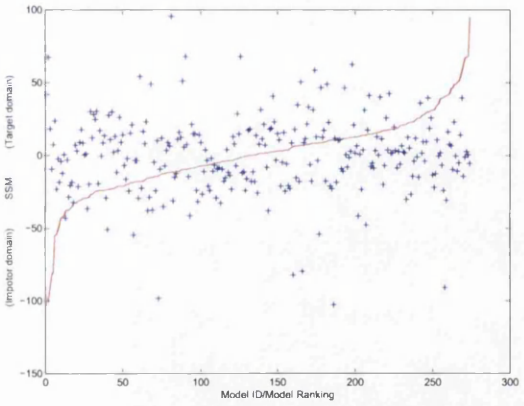


Figure 5.2: SSM measure given per target model, illustrated by the x-axis are trained on 10sec utterances. The y-axis represents the derived SSM score. The ranked target SSM is shown by the red plot

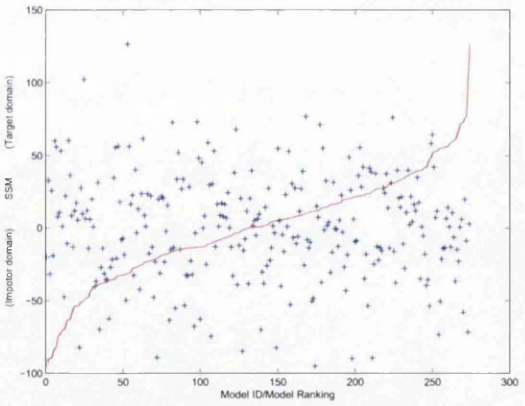


Figure 5.3: SSM measure derived from the 1conv trained targets, identified on the x-axis. The y-axis represents the derived SSM score. The ranking of these scores is depicted by the red curve

male, 10sec targets from the NIST 2005 evaluation and 274 male 1conv targets in Figure 5.3. Each target, depicted by the blue '+' markers, is scored against approximately 1000 impostor utterances. The x-axis depicts the target model ID and the y-axis relates to a the SSM quality observation. High SSM measurements are thought to give an indication of a poor model. For the 1conv models from the same evaluation database, a few targets can be seen to show low scores (bottom-right of Figure 5.3), indicating potentially false acceptance errors to assumed impostors. A majority of models surrounds the zero score. SSM's above zero can be deemed more resilient to impostor

utterances. The targets perception of a true trial is still unknown, though this could of course be established with an adequate supply of target utterances. These averaged scores are ranked by order of their SSM, shown as the red plot in Figures 5.2 and 5.3, high to low to give us an indication of the robustness of a target model against a volume of impostors. A high rank (e.g. rank of 1) relates to a high SSM score.

Target models which reside in the low SSM domain and have only been assessed against impostor utterances. These can be deemed *of poor similarity* in real verification trials and could either be rejected immediately or weighted accordingly; similar to approaches described in [47, 9]. Derivation of an enhance weighting procedure could be an avenue of investigation; One hypothesis could be to derive optimum weighting parameters through some calibration procedure by utilising a data-driven approach with a large database of trials. Female targets trained on the same condition display similar distributions characteristics to the male population.

Koolwaaij et al. [45] reports that quality can only be derived from the behaviour of a model not on the model directly. A measure on the model directly would be difficult without a pre-defined reference scale of model quality. A simple yet coarse quality estimate can of course relate to the duration of target speech and hence the number of feature vectors used during enrolment. Figure 5.4 presents the 10sec target SSM against the number of feature vectors used to train the target. The 1conv target training scenario is depicted in Figure 5.5. When applying the SSM approach described in this chapter, we can see in the 10sec task, that models of low feature vector content are ranked low, hence *poor* quality. One logical observation of a speaker-dependence model quality

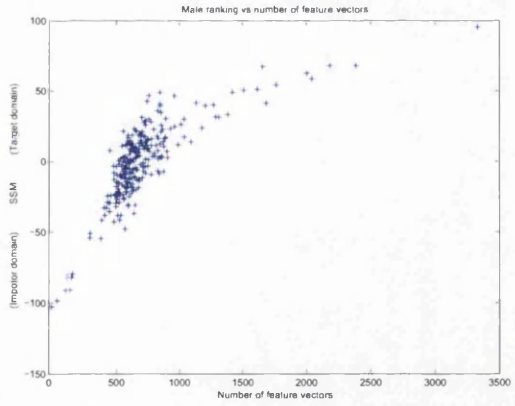


Figure 5.4: Number of features (x-axis) vs. SSM score (y-axis) for the male and female target models trained with 10sec utterances

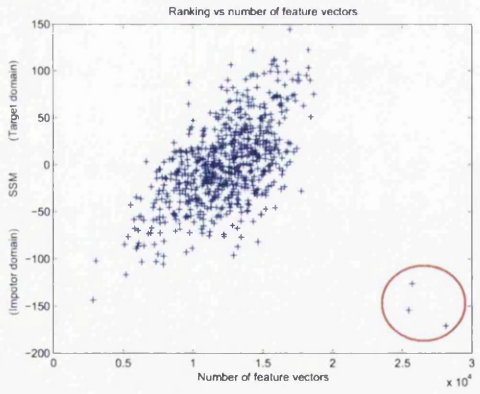


Figure 5.5: Number of features (x-axis) vs. SSM score (y-axis) for the male and female target models trained with 1conv utterances. Distinctive, outlying *goats* are highlighted with a red circle

can be the duration of speech used to enrol a speaker, usually governed by the number of feature vectors used, assuming that the content of the feature vectors is useful speaker-dependent informa-

tion. A negative correlation can be seen for both 1conv and 10sec tasks, where models trained with a quantity as low as 21 vectors show low speaker-discrimination characteristics through the SSM approach. A speaker model of low speaker-discrimination is of course going to show a similarity to other speakers, hence the low SSM score to flag under trained speaker models.

For the 1conv models, a less defined negative correlation can be seen. Target models can be seen to reside outside the general score range when the SSM and feature vector norm are shown in the bottom right-hand corner in Figure 5.5. These *poor* models contained high amounts of feature vectors but with low mean quality scores indicating a *goat*. Only a minor subset of models from both genders were found to reside in this outlying region and three models can be seen to do so in Figure 5.5. Manual analysis of these speakers revealed a content of approximately 20 seconds of speech in the conversation with silence manifesting the remaining 5 minutes of the utterance. Application of the CMS normalisation together with the speech non-speech detection threshold has allowed a large quantity of silence features (approx. 25000) for use in speaker adaptation.

AT-Norm would possibly remove impostor models that display *goat* attributes from a general targets normalisation cohort. However, targets that display *goat* characteristics would likely include equivalent *goat* impostor models into their personal impostor cohort.

5.3 A subjective assessment of target quality

An examination through subjective listening tests of target utterances found to exhibit unusually high likeness to impostors through the SSM approach is now described. Attempting to reveal some insight to account for a targets high false acceptance rate.
SSM score observations from the 10sec models taken from both of the SSM ranking extremes, shown on Figure 5.2 are manually characterised and presented in Table 5.1. Few models from the centre of the distribution are also assessed on reference utterances expected in the 10sec task.

The first column in Table 5.1 indicates the utterance used to generate a target model. The derived SSM rank from the target subset is given in the second column, where a high ranking measure corresponds to model trained with low quantities of feature vectors; an utterance containing 21 features provides the lowest SSM score. The third column depicts the number of feature vectors obtained from the utterance for model training. These models are divided into three subsets, categorised through manual subjective assessment. *Poor* utterances contain far less the expected 10 seconds of speech and such low speech duration and subjectively does not appear to have distinguishable speaker content. Comments to the particular utterances from the subjective listening tests are given in the last column. It is obvious that as the duration of speech increases, the utterances and hence the trained models can be labeled *good*. Models trained with around 2000 feature vectors give high SSM scores and hence a greater rank, indicating the most well trained models from this target cohort. The content of speech applied to these models in general had a variety of speech context During a trial, it is expected that such poorly trained models, e.g. jecv.B would always produce low scores, however this is not the case as enrolled model bears resemblance to other speakers, i.e. impostors, hence high likelihoods when scored against other speakers.

Utterance ID	SSM Rank from measure	Number of feature vectors	Subjective quality	Comment
jims.B.sph	1	3329	Good (Sheep)	Lot of '*yeah*', not much variety, low background noise. In total, over 1 minute of file, approx. 10sec of speech
jbrf.A.sph	3	2384		Variety of speech
jdik.sph	2	2180		High SNR with speech, with a little laughter, little under 10sec of speech
jeua.B.sph	162	659	Expected (Sheep)	Variety of speech with repeated 'Oh'
jeni.B.sph	42	659		Variety of speech context
jhqi.B.sph	93	657		Different language, lots of variation of speech, just under 10sec of speech content
jgrm.B.sph	271	145	Poor (Goats)	Speech though just over a second of content
jihr.A.sph	272	132		Speech though stuttering and incomplete, total 3.7 sec long with silence
jgcq.A.sph	273	64		Background noise, 1.4 sec long
jecv.B.sph	274	21		No distinguishable speech, 0.55 sec utterance

Table 5.1: Subjective assessment of key target models from the 10sec evaluation of NIST 2005. Each labelled utterance is profiled against the derived SSM score, with high rank beginning at 1. Also the number of feature vectors used to generate the speaker specific model, general subjective assessment and utterance comments are provided for each utterance. It can be seen that utterances of low SSM rank contain little speech, confirmed by a blatant lack of speech features to enrol the speaker.

Thus far, it seems that the content of speech within an utterance is related to the number of feature vectors extracted. However, as seen in the 1conv scenario, few models trained with relatively large quantities of feature vectors still portray poor models by providing high SSM scores. The range of trial scores from the NIST 2005 evaluation can be seen in the score distribution, generated for both 10sec and 1conv tasks in Figure 5.6 and 5.7 . Without appropriate quality detection, the threshold may need to be set higher to counter the score deliverance from the poor models.

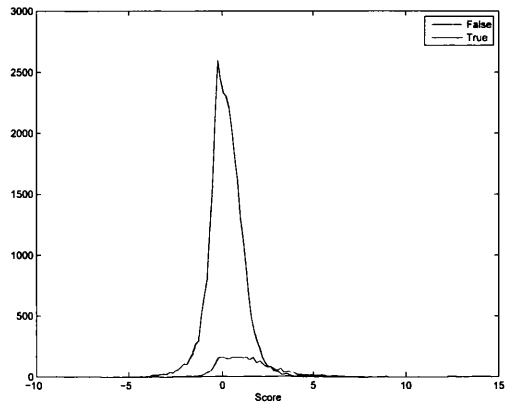


Figure 5.6: 10sec true and false score distribution. Score plotted on the x-axis and cumulative quantity of models depicted on the y-axis. Notice different y-axis cumulative range.

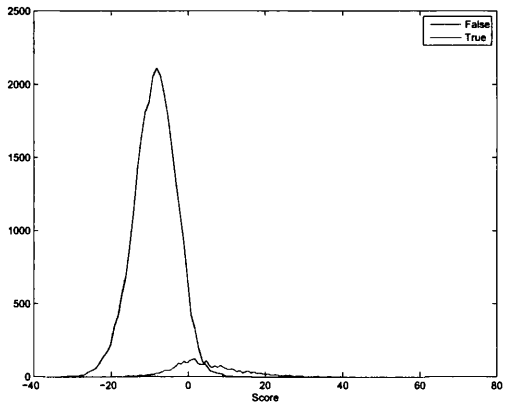


Figure 5.7: 1conv true and false score distribution. Score plotted on the x-axis and cumulative quantity of models depicted on the y-axis. Notice different y-axis cumulative range.

A poorly trained model, enrolled with utterance *jecv.B* are provides scores between 0.45 and 1.16, all of which are scores from impostor trials, score normalised with T-Norm. It was found that model *jeni.B*, trained with a duration of 10sec speech, delivered between 0.12 and 1.25 for true trials and between 0.02 and 1.0 for the false trials. Thus a high threshold would be required to overcome the large overlap of trial scores given by the poor and well trained model. Model quality measures could be preliminary used to filter speakers that generate such overlap which are known to be poorly enrolled. For the 10sec target models, computational overhead to apply the quality measure approach can be avoided with common sense logic to dictate the viability of a target model, e.g. models trained with only 21 vectors would unlikely be able to represent a specific speaker and could be acted on appropriately prior to some measure of quality.

During the score stage of a trial, an alternative approach to a model weighting or dismissal procedure could be to employ a recursive feedback loop to tune the speech detector, allowing more speech data through to enrolment. Though in the case of utterance *jecv.B* (an utterance found to contain very little speech), applying such a tuning procedure could further corrupt the speaker model by

allowing more non-speech distortion into the model. This approach could result in degraded models residing in the outlying region, shown in Figure 5.5.

In Figure 5.5, the SSM has highlighted 1conv models that could potentially provide many false acceptances, even though many feature vectors are used for training. The SSM is particularly suited to highlighting the general speaker-related content of the training utterance and hence the validity of a target model. Errors filtered from the speech detection stage and the feature level have been identified through the speaker model and could be appropriately compensated.

5.4 Experimental Observations with the SSM

This section investigates the potential of the SSM through a simple confidence-weighting of trial scores from certain models, an approach similarly applied by Koolwaaij et al. [45]. The SSM is applied to the target models from the 10sec and 1conv tasks in the 2005 NIST evaluation. For confidence in the proposed model quality measure, the same protocols are applied to the same conditions of the NIST 2006 evaluations.

Here, a simple exponential weighting function is used to perform the score biasing. Exponents from 1 to 10 were investigated, revealing that an exponent with an power of 3 delivered best performance. The gain of EER and NIST DCF is minimal, 0.03% and 0.005 respectively, though the false alarm region error rate can be seen to have improved in both 10sec and 1conv evaluations. This results is expected as the approach highlights models that cause high amounts of false alarms. For the

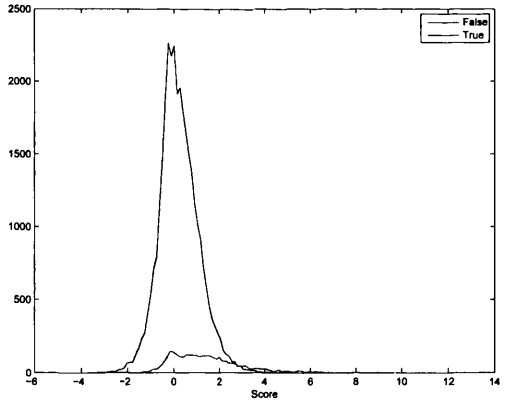


Figure 5.8: 10sec true and false score distribution with applied quality weighting. Notice different y-axis cumulative range to system with no SSM pre-filter in Figure 5.6.

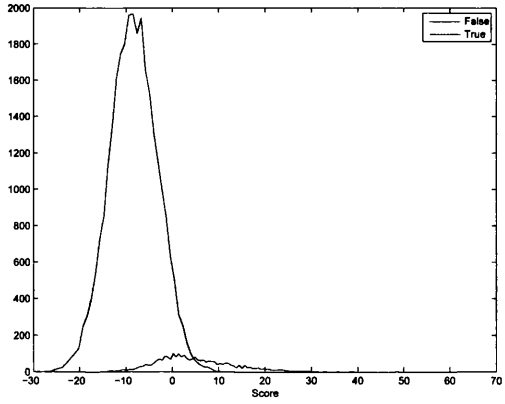


Figure 5.9: 1conv true and false score distribution with applied quality weighting. Notice different y-axis cumulative range to system with no SSM pre-filter in Figure 5.7.

10sec task, by contrasting the conventional score distribution in Figure 5.6 and the SSM weighted scores illustrated in Figure 5.9, a decrease in the overall amplitude of the false distribution can be seem when the SSM procedure has been applied. The variance of the distribution has increased

slightly with a mean shift, though no divergence of means is shown between the true and false distributions. In Figures 5.8 and 5.9, two notches can be seen around the zero score in both the 10sec and 1conv false distributions when the SSM has been applied. However, the results display

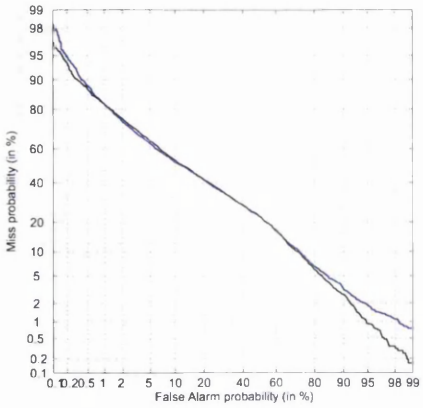


Figure 5.10: 10sec NIST 2005 evaluation performance with applied quality weighting. Blue represents conventional T-Norm, the black profile includes SSM

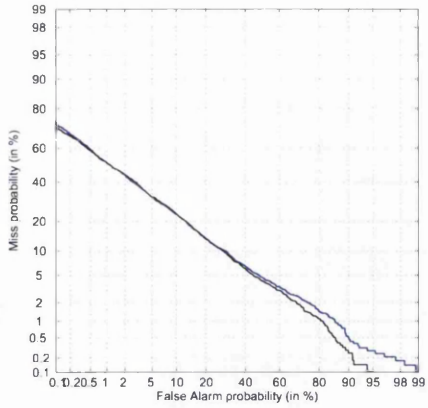


Figure 5.11: 1conv NIST 2005 evaluation performance with applied quality weighting. Blue represents conventional T-Norm, the black profile includes SSM

the use of sub-optimal weighting approaches, though this does illustrated the potential effectiveness of the SSM model quality information when applied in the score domain.

This approach has two benefits, with the abundance of impostor examples, where scores can be simply generated. Secondly, the measures can be derived off-line in training from any classifier that can derive a verification score between two utterances. However, during a scenario where continuous speaker adaptation is applied after a successful verification, the SSM observations would require re-computation. Such model quality observations are palatable for surveillance application, however, this is not viable in the NIST cost domain. Generating SSM observations on target models scored against a collection of target specific utterances would likely provide greater enhancement in the security application domain (region 1) concentrated on by the NIST speaker recognition evaluations.

To display confidence in this approach, the SSM is also applied to the same task conditions in the NIST 2006 data set. The performance of the 10sec task from the NIST 2006 evaluations, illustrated in Figure 5.14 do not benefit from the weighting of the SSM. Target model SSM scores shown in Figure 5.15 do not deviate. Included on Figure 5.14 is the conventional T-Norm measure without the mean component. This normalisation approach, from results for the 10sec task presented in Chapter 3 usually provide some distinguishable enhancement, though here, only a minor EER improvement can be seen. The 1conv task from the NIST 2006 evaluation presents a slight improvement with the addition of the SSM process, similar to the false alarm enhancement in the NIST 2005 scenario. A 60% relative improvement can be seen on the Miss probability when considering high false acceptances. This of course is not a rational operating point increase as

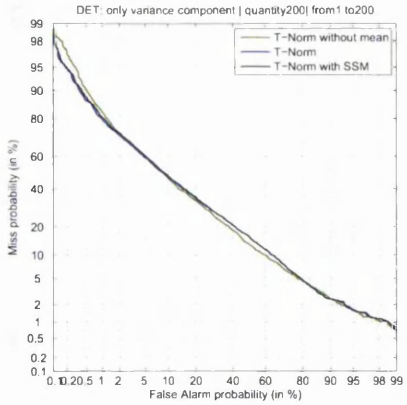


Figure 5.12: DET performance plot showing 10sec task performance with NIST 2006 evaluation.

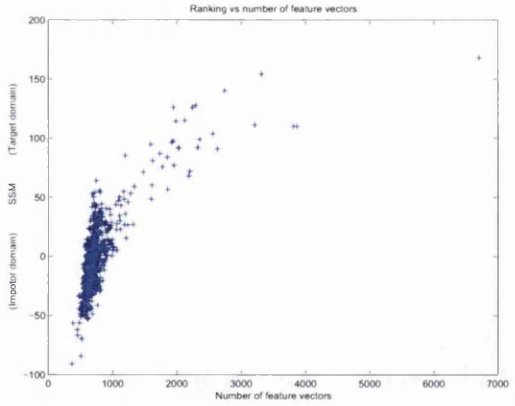


Figure 5.13: Number of feature vectors (x-axis) vs. SSM scores (y-axis) to derive the speaker with the 10sec task from NIST 2006.

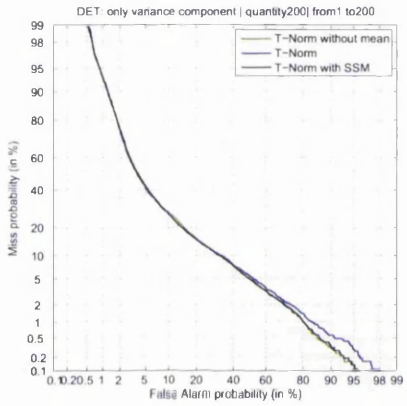


Figure 5.14: DET performance plot showing 1conv NIST 2006 evaluation performance with applied quality weighting.

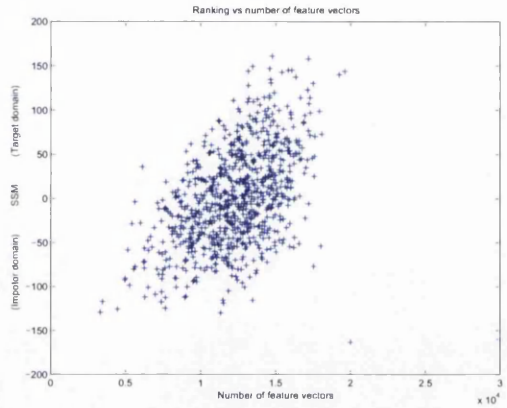


Figure 5.15: Number of feature vectors (x-axis) vs. SSM scores (y-axis) for each target from the 1conv NIST 2006 training set.

only around 10% of impostors are discarded. These NIST 2006 results support our findings on the 2005 development set when poor target models are present. The potential for discarding extreme circumstances of poor models has been shown. From SSM evaluations discussed here, only 0.5% to 1% of such models exist and in the NIST 2006 10sec-10sec task. Further experiments are required to examine the potential for such a model quality measure with optimum weighting parameters.

A recent hypothesis from the speaker verification community expresses concern over the viability of the NIST 2006 evaluations. Trials have been shown to exhibit labelling inconsistencies between genders, where impostor trials have been wrongly labeled as male speakers when in fact they are female speakers, deriving lower scores and hence better performance.

5.5 Discussion

These experimental results have determined if the produced trial scores can give further information and subsequent confidence from a speaker utterance. Some speaker models are derived from utterances that can be degraded or overwhelmed with non-speaker information, subsequently providing poorly adapted models. Such outlying target models can be a source of decision error. The potential for the SSM model quality observation to counter such models has been examined in this chapter. The SSM approach can be a convenient way of predicting extreme false acceptance errors supplied by poorly adapted target models. Observations are gathered from a collection of tailored false alarm trials as the SSM data-driven approach has, in this scenario, an abundance of impostor utterances. A collection of target utterances could also be applied to ascertain a targets distinctive nature among other speakers. Of course, an approach to avoid such poor models would be to simply discard such models that fall below a pre-defined threshold. However, analysis of the SSM shows that some erroneous utterances of expected feature vector quantity can exhibit poor characteristics. Here, the SSM scores have been derived from the GMM classifier, though any classifier that generates a score between two utterances could be used to generate SSM observations.

In this approach, the enrolled speaker confidence is analysed through their derived model, however, the test utterance has not been examined which could also give errors. This could be overcome by conducting bilateral scoring along with a parallel SSM approach. Bilateral scoring is a reversal of roles of both trial utterances where the test utterance is employed in model enrolment and the training utterance is considered as the test utterance. Further work is necessary with this approach to ascertain the confidence in a test utterance. The D-Norm approach, discussed in section 2.7 may assist with models trained on extremes of utterances for a given task, potentially providing a non data-driven approach for deriving a model quality measure.

To ascertain the viability of the SSM, a simple weighting approach is applied to illustrate the potential score compensation of poorly derived target models. Performance enhancement is negligible in the NIST cost domain and the EER in all short and longer utterance tasks. However, there is an increase of performance in the low Miss probability domain where lower score thresholds are set for surveillance type applications. The SSM has been shown to be a potential pre-filtering approach that can be deemed for further examination.

Chapter 6

Conclusion and Future Work

In this chapter, points discussed throughout this thesis and final thoughts are brought together with an overview of potential avenues for further work.

6.1 Conclusion and discussion

Normalisation is key to performance in speaker verification (SV), attempting to reduce the effects of degrative signal perturbations and speaker differences. Approaches such as test-normalisation (T-Norm) have become commonplace in SV to reduce signal perturbations and was used by over half of the competitors in the most recent international speaker recognition trials conducted by NIST. This work has focused towards understating the impostor-centric observations and their ramifications on SV, primarily, when the popular score normalisation approach T-Norm is applied. Selection of the normalising cohort in the score domain is known to enhanced performance and this thesis has been primarily attempting to exploit the use of prior knowledge through enhance selection. Due to the availability of impostor utterances, the impostor-centric route is applied to gather normalisation statistics. However, as discussed in Chapter 2.3 a target-centric approach may produce extra informative normalisation parameters.

The trial-independent cohort selection scenario was discussed in Chapter 3 and was found to give greater enhancement for a given trial when composing a cohort of similar characteristics to that of general set of concerned target speakers. Here we have described the different impact of selection under different task conditions where shorter durations are prone to higher variability in performance without some form of impostor selection. Impostor cohort compositions with random selection procedures illustrated a high variability of system performance. However, exploiting some prior-knowledge in the form of a target's average training duration allowed the matching of the impostor cohort with a reduction in decision errors. However, deliberate misuse of prior knowledge can lead to a poor selection of impostors found to degrade robustness by approximately 8% at the EER in the 10sec-10sec task. Several impostor cohorts contained intentional assortments, with a

majority of either matched or miss-matched were engineered to illustrate the large variability and danger of selection, especially in the 10sec-10sec task. Here we have shown that when applying such selection approaches in a trial-independent manner, as few as 15 impostors can be used in 10sec evaluations providing admirable results. With 1conv utterance enrolment, it was found that a cohort containing in the region of 25 impostor models could be utilised for T-Norm.

Further investigation led to the dissection of the T-Norm approach, observing that the standard deviation only approach rather than the normal mean and standard deviation, generally gave superior performance. From empirical analysis, it is observed that the mean component suffered from high sensitivity with certain combinations of targets and impostors with increased verification errors. This effect is reduced when the mean component was removed from the normalisation procedure. This was found to be a contribution from a small contingent of poorly enrolled target speakers that delivered abnormal scores when normalised against the T-Norm mean (highlighted through the SSM approach). Removing these then renders the conventional T-Norm (including mean) the appropriate approach.

Trial-dependent cohorts can further reduce verification errors by supplying each target with a specific cohort of impostors, selected during the enrolment process. Such cohorts can be generated through the adaptive T-Norm (AT-Norm) approach. This can alleviate the effects of poorly derived target models by normalising against a personal cohort of similarly characterised impostors, e.g. normalising a poor target using poorly trained impostors. AT-Norm provides better performance and enhanced computation efficiency at test time by further reducing the number of impostors to supply a personal normalising distribution. A disadvantage lies at training time to gather score observations to performing the selection procedure through testing. As this can be performed effectively off-line, real-time operation is not hindered. The pool must also be large enough with a variety of impostor attributes to accommodate a decent selection approach. AT-Norm has been shown to give superior performance over a trial-independent approach in all trial conditions.

As shown in this thesis, scores can be utilised for many aspects of the classifier, from discrimination of target models to a data-driven approach for setting a decision threshold for subsequent trials. Observations from the T-Norm approach led towards the *speaker security measure* (SSM) discussed in Chapter 5. Essentially, the statistics gathered by the impostor cohort for T-Norm is hypothesised to describe further informative measures of a speaker's viewpoint for a given speaker over a collection of impostor attacks. SSM observations can describe the confidence of an enrolled model against a large collection of impostor attempts. The SSM approach can reveal, on rare occasions, speaker models trained with a high proportion of feature vectors with little speaker-specific content, likely to be prone to false acceptance errors. This has shown to be of possible benefit in evaluations where poorly trained speaker models are present.

The correlation between the number of feature vectors in an utterance and the confidence of the speaker model can usually be a good indicator for a subsequent score weighting regime. The SSM can be applied in tandem to enrolment for highlighting potentially troublesome target speakers. Preliminary results show promise to detecting such poorly derived models. From observations of the SSM investigation, it was found that this approach coupled with conventional T-Norm performed similarly to the standard deviation only T-Norm procedure discussed in Chapter 3.6. Where the SSM appeared to remove the degrading effect of the mean component. Further investigation highlighted the few poorly enrolled target models that contributed to the mean degradation during the evaluations. However, conclusions drawn for the proposed SSM approach must be taken with care since only a preliminary investigation has been conducted on observing the resilient influence toward impostor attempts.

6.2 Further work

Two themes have been presented in this work, foremost of which is score normalisation with analysis concentrating on the T-Norm approach. A preliminary investigation of target model confidence has also been discussed. Strategies towards further work on approaches discussed in this thesis are presented here.

Many combinations of the approaches discussed in this thesis could be performed. One example for further investigation is bilateral scoring; this can be applied with score or decision fusion. The integration of other useful *prior knowledge* towards model selection for T-Norm is another avenue to consider.

Further normalised score observations could also be collected from other classifiers. For example, the popular SVM with GMM supervectors could be used, providing the benefit of pre-normalised scores for selection. This can of course be performed with any classifier that produces a score based on an observation. However, it is difficult to predict the usefulness of sequential normalisation methods such as T-Norm in the score domain when some utterance degradation is combatted at earlier stages. The selection procedures that utilise prior information could also be transposed to other forms of normalisation, for example, a speaker specific impostor cohort for Z-Norm. Approaches developed during this research can now compensate for utterance miss-match during earlier stages of the classifier, e.g. in the modelling or speaker enrolment domain. There are several areas within the classification system where compensation to signal perturbations can be applied. Gravier et al. [4] shows experimental results with the introduction of prior knowledge at different stages of the classification system. They concluded that a system benefits performance by introducing prior

knowledge early in the classification system. It is observed that current state-of-the-art trends seem to support this by moving prior information toward the beginning of the classification chain, highlighted by Kenny et al. [15] & Fauve et al. [24]. The ramifications of adaptive features from observations by Fauve et al. [24] coupled with different sequential score normalisation could also be considered.

The introduction of the SSM approach has shown advantageous results by weighting target models of low speaker discrimination. Simple weighting approaches presented in this thesis could be superseded with other functions which could possibly provide better model quality demarcation. On a note of computational efficiency, the optimum number of features or data-drivers to provide a confidence in the SSM scores needs to be considered. The precise nature of the SSM has yet to be fully understood and suggests further investigation.
The recent work by Ferrer et al. [20] in March 2008 leads to a different strategy of introducing SSM type information through a score fusion manner. This could be achieved with the aid of the FoCal toolkit, developed by Brummer et al.[1].

The statistics gathered by the impostor cohort for T-Norm usually contain a variety of informative measures, scores of which can be considered as utterance coordinates among a cohort of assumed impostor speakers. This approach is rising in popularity and is commonly termed *anchor models* [48], where a speaker is represented by a cohort in a speaker related reference space.

These observations lead to many avenues to future work.

[1]FoCal Bilinear Toolkit available at http://niko.brummer.googlepages.com/focalbilinear, as of March 2008

Chapter 7

Epilogue

In late March 2008, NIST released an updated 2006 speaker recognition evaluation key, describing a list of trials with associated ground-truths that were deemed 'bad' and should be removed. As the final part of the experimental work presented in this thesis, results using this new key are presented.

Despite of trial errors in the original NIST 2006 results, a large leap in performance is evident from a variety of submitted systems from previous NIST evaluations. This illustrates the ability of state-of-the-art systems by demonstrating a high robustness against a sub-set of null trials.

7.1 Invalid Trials of the 2006 NIST Evaluation

Throughout this work, labelling of the speech utterances provided by NIST, in terms of utterance identity with ground-truth (primary) and prior knowledge (secondary, for example, gender), is assumed to be correct. However, in any large database like these of NIST, errors are likely to exist. In practice, such primary labelling errors would be difficult to detect and can impact on true system performance. Such primary utterance labelling errors can come about in many ways, one of which is when an individual wishes to claim two identities during the period of collecting speech. This could be a genuine mistake or produced with malignant intent, one possible hypothesis is financial gain. If this were the case, it can be postulated that a trial of a target model and test utterance are deemed as a false trial, when in truth, both target and test utterance, although labelled as different speakers are in fact the same person. In this scenario, an incorrect false decision is likely. A less serious case of deliberate error in the secondary case could be miss-labelling of the handset if it, for example, has been changed over a period of speech collection. Any other secondary labels also falls into this category. The use of *pre-filters* to highlight such errors become important. These errors are more sensitive to well tuned systems, e.g. statistical data-driven approaches that are dependent on a criterion of data as background knowledge for classifier construction. Usually, evaluations are likely to overestimate the accuracy of a data-driven classifier due to database tuning. This is usually comes

from assembling a classifier from development sets which contain common characteristics across the databases. However, with labelling errors in the original NIST 2006 database, the evaluation leads to a degradation in performance and the revised keys lead to improved performance.

Condition	Gender	Number of bad trials	Total number of trials	% of trials removed
10sec-10sec	Male	511	15013	3.29%
	Female	563	18540	2.76%
1conv-1conv	Male	1194	23292	5.02%
	Female	1786	30673	5.64%

Table 7.1: This table illustrates the reduction in trials between the NIST 2006 new and old key, bad trials refer to the removed trials.

Table 7.1 presents a break down of invalid and missing trials from the revised NIST 2006 ground-truth key. This illustrates an approximate total reduction in the number of trials for the 10sec-10sec and 1conv-1conv evaluation is 3% and 5% respectively. This is a high proportion of trials to have been subject to poor labelling. The DCF and EER performances are given in Table 7.2 for both 10sec-10sec and 1conv-1conv NIST 2006 evaluations. The total number of trials using the revised key is also provided. With the removal of such trials, the 10sec-10sec NIST 2006 evaluation performance, illustrated in figure 7.1, has little change. Although, the removal of ambiguous trials from the 1conv-1conv evaluation, depicted in figure 7.2, shows a very different trend. Here, the performance has increased significantly at the NIST DCF operation point from 0.087 to 0.062 with an approximate 1.5% EER reduction when a total of 2980 male and female trials are removed. The sharp decrease of performance in high acceptance region (region 1) of figure 7.2 is a result of the invalid database labels which has been increase significantly with the new key. It is interesting to observe, that by chance, the gain provided in the 1conv-1conv evaluation with the removal of the bad trials is not shown in the 10sec-10sec task. This is based on the trial configuration derived by NIST.

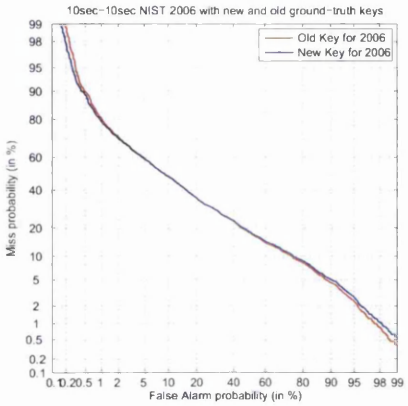


Figure 7.1: DET performance plot showing the 10sec-10sc NIST 2006 evaluation performance with revised and original decision keys.

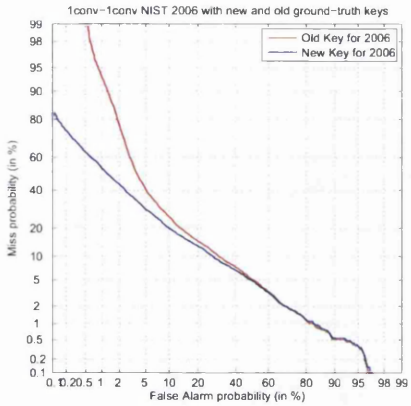


Figure 7.2: Performance of the 1conv-1conv NIST 2006 evaluation performance with revised and original decision keys, represented by a DET plot.

Condition	Key source	% EER	DCF	Total number of trials
10sec-10sec	Old	29.16	0.0896	33553
	New	29.18	0.0886	32550
1conv-1conv	Old	17.08	0.087	53966
	New	15.54	0.062	51068

Table 7.2: This table illustrates the performance at the DCF and EER with the revised and original keys for the 10sec-10sec and 1conv-1conv NIST 2006 evaluations

7.2 Final Thoughts

Over the period of the work presented in this thesis, technology has improved by such an extent, the accuracy of database labelling has been drawn into question. This comes to the fore in system development through error analysis. This issue has been highlighted here when using the SSM to detect extreme miss-labeled utterances in terms of expected utterance duration. This illustrates the difficulty and importance of large representative databases. As the performance of systems improves, database labelling errors become more significant. System error rates fall with system improvements, hence miss labeled errors become a more significant percentage of the total errors and more easily identified. The lower numbers makes label checking more viable, for example, when errors rates fall below 10%, the number of human-based checks required reduces to x. However, this is only one class of error, where the two utterances from the same speaker, labeled as different people has been deemed false through a trial, when truly, they are the same person. To avoid such bad trials, the labelling of the utterances is critical. The derivation of the revised ground-truth key reflects advancements in technology, from a standpoint of being able to question the database. The state-of-the-art systems have helped to highlight a number of labelling errors.

Appendix A

Appendix

A.1 Classifier configuration

Research and development is primarily conduced on the NIST 2005 and 2006 evaluations. The NIST 2004 dataset will be used to represent the source of 'other' prior knowledge and is also used to build both UBM and normalisation cohorts. Specifically, the normalisation cohorts are gathered from the training utterances of the dataset.

Features are extracted in the same manner from both the claimant and the target speaker in which the target features are used in the modelling process. A single vector constitutes of a window of speech, typically 20-30 ms in duration. By sliding the window along the speech signal with 50% overlap, a series of vectors are extracted using cepstral derived coefficients from a Mel-scale filter bank, producing a chronological sequence of feature vectors. Traditionally in the GMM, the features of a speech sample are quantized into a vector of $1 * 16$, where a $16\hat{t}h$ order feature vector represents a windowed section of the speech spectra. Appended to the 16 baseline extracted features are 16 delta components the baseline energy and delta energy coefficients. A post-process to removing the silent elements is conducted with a tri-Gaussian speech detector with a zero mean, unity variance normalised distributions. The alpha threshold component of the speech detector is set to zero. All features are further processed to remove linear utterance characteristics through cepstral mean subtraction (CMS). No further conditioning or gender dependent optimisation's have been applied to the features.

The GMM classifier is used with 1024 model components trained on the test set of the NIST 2004 data set. The initial background codebook is generated by applying VQ, specifically through the Linde-Buzo-Gray (LBG) algorithm. Approximately 1000 feature utterances, each containing an average speech duration of 2.5 minutes. The expectation maximisation (EM) algorithm (highlighted in [19]) to 10 iterations is used to monotonically increase the likelihood of the codebook until the training data applied converges. Speaker-specific modelling is applied with mean only Maximum A Posteriori (MAP) [18, 19] adaptation with a fixed relevance factor of 16. Utterance scoring

is conduced using a fast-scoring routine, indexing the closest 5 ranked components for each trial. Score normalisation T-Norm impostor cohorts have been acquired from the NIST 2004 database. The default pool of 600 models contains 3 sets of 200 models based on the utterance duration. 200 are trained on 10 second speech duration (abbreviated to 10sec), 30 seconds of speech (30sec) and approximately 2.5 minutes of speech (1side). The composition of the cohort is modified accordingly during the course of this thesis.

A.2 NIST evaluation database

The National Institute for Standards Technology (NIST) conducts annual speaker recognition evaluations, allowing world class institutions to compete using large-scale databases over different speech conditions. Primarily, conditions are combinations of utterance durations and handset capture devices in both training and testing data. Each condition ideally has a different decision threshold and is usually set by conducting development trials on previous years with the supplied ground truth. Also note that, in accordance to the NIST protocol, all trials are performed independently where no unsupervised adaptation has been conducted between trials. The evaluation plans for the NIST speaker recognition trials for years 2004, 2005 & 2006 can be found at [49]. The 1conv-1conv is the required task to be undertaken for system submission. There are combinations of different speaker trials, defined by duration for both training and test data. These conditions have slightly varied over the three previous years. The 10 and 30 second utterances are excerpts from a selection of 1conv conversations. The 1conv utterances consist of approximately 5 minutes of a conversation from both parties with silence replacing the speech of one speaker with roughly 50% speech from each speaker. One goal with this scenario, though not considered here, is to separate the individual speakers during verification. Utterances are split into genders and verification between two utterances are likely to use different handsets. The primary conditions chosen for experimentation are the 10 second training and 10 second testing conditions, denoted as 10sec-10sec and the 1conv-1conv condition. The hyphen separates the condition of the trial to training-test type.

The formality of the data for a standard Speaker Verification (SV) evaluation shall now be discussed. The speaker data is split into two categories, a development and evaluation set. The utterances contained within each set is assumed independent from each other. The development set allows for closed loop system enhancement through experimentation. For a true evaluation, the evaluation data must be deemed as unobserved data, i.e. has no ground-truth from which we can draw conclusions or prior knowledge. This satisfies the application scenario of a real trial where no ground-truth is available to validate the decision provided by the ASR. No feedback is supplied from the NIST 2006 database. To generate the experimental evaluations for these experiments, the ground-truth is available, though this is only used to generate the true and false distributions

to illustrate system performance. The experimental results show in this thesis are collected using data sets from the NIST speaker recognition evaluation from years 2004 & 2005. The 2006 NIST database will be used to validate the results.

All attributes of the system are to be considered as default, unless otherwise stated.

A.3 Understanding results as a function of the error rate

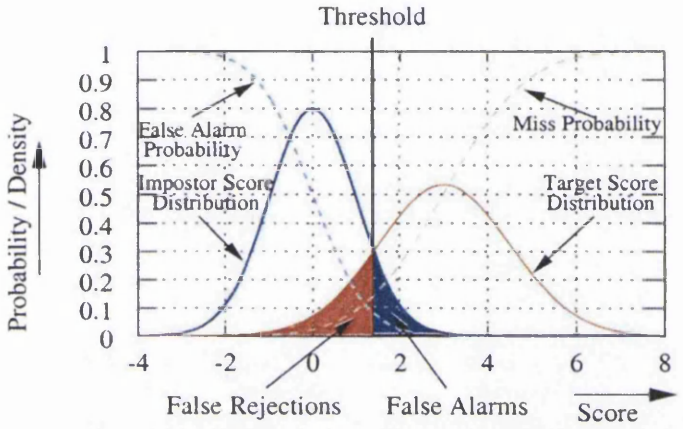


Figure A.1: Example of true and false score distributions from an evaluation

Figure A.1 depicts an example of the and distributions from a group of trials. Verification uses error rates to determine the performance of a system. Two types of errors can occur, False Acceptance (FA) and False Rejection (FR). Possible verification outcomes are shown in Table A.1, red denotes an error in verification whilst green states a successful verification. These errors are the metric used to judge performance of a speaker classifier. FA consists of accepting wrongfully, a claim by an impostor. The latter wrongfully rejects a valid user. Both of these errors are used to set a threshold. Setting the threshold to low will accept many impostors therefore having a high FA. Setting the threshold high, and the true speaker will have a higher rejection rate. This is a balance of application based influence; a convenience (low FR) and security (high FA), both having different uses in various applications therefore making the threshold setting an important factor when deploying as an application. Again this is an application specific threshold setting. Practically the threshold is usually set using a development corpus (where the trial ground-truth is known) by logging the frequency of each error. Large development sets are required to give confident performance statistics. Detection Error Trade-off (DET) curves is another means to represent the error statistics as graphical representations on a 'normal deviate scale' [41]. In the Figure A.1, the two types of errors, FA and FR are shown as the filled regions or blue and red respectively. To

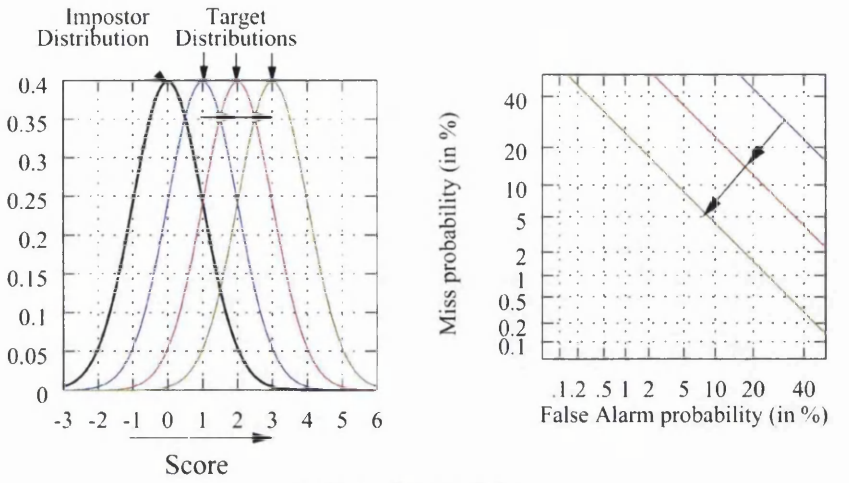
obtain a DET curve we plot the FA vs. FR on the normal deviate scale. An example of ideal DET curves are shown in Figure A.2, courtesy of Roland Auckenthaeler.

	Accept	Reject
True speaker	Acceptance Probability	Missed Probability
Impostor	False Alarm	True alarm/False Acceptance

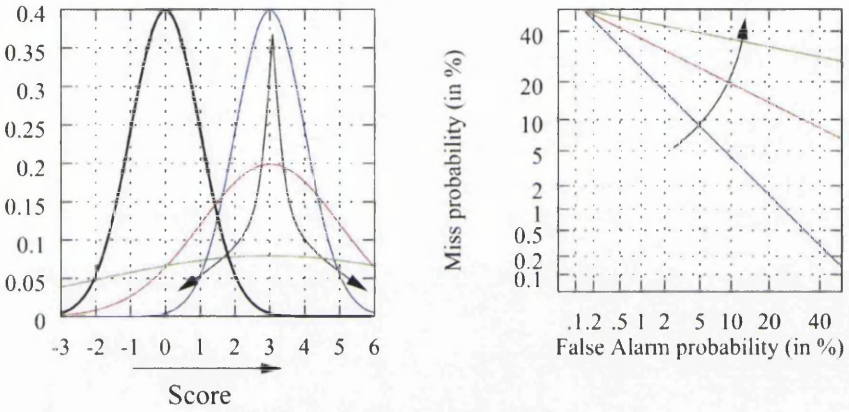
Table A.1: Interpreting the outcome of a trial

SV performance can be visualised through the use of the detection error trade-off (DET) [41] curves, allowing observations on a systems variation and assist to set an applicable threshold. Navrati and Ramaswamy [11] show that the more Gaussian the true and false distributions display, the greater linearity of the DET curve. This is also ideally explained by Alvin et al. [41]. The *tilt* observed by Auckenthaeler et al. [10] is shown by the greater variance constriction of the false distribution when T-Norm is applied, illustrated in Figure A.2 (b) and (c).

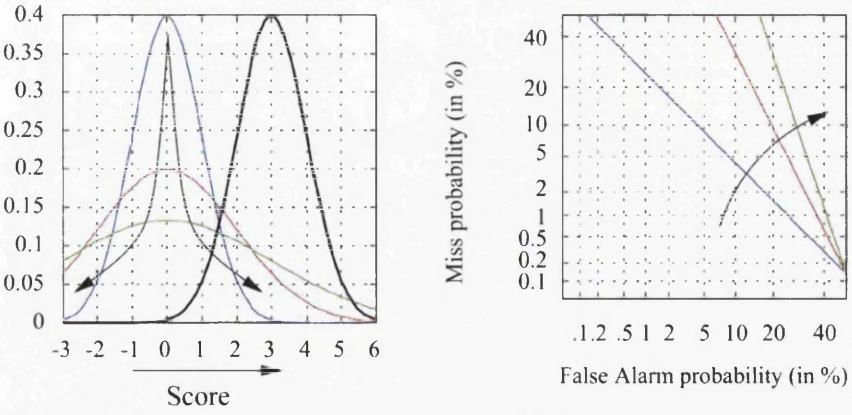
By synthetically modifying attributes of the true and false distributions, its effect in the DET performance measures shown in Figure A.2. This illustration helps identify the changes in the DET plot, provided for the convenience and to assist to the reader. The true (target) distribution is usually considered a positive score offset to the false (impostor) distribution. Plot (a) describes the fundamental goal to provide robust verification by separating out the resulting true and false distributions from an evaluation with equivalent consequence in the opposite DET plot. The outcome of scaling the true and false distributions show a pivoting of the DET curve, shown respectively by sub-plots (b) and (c).



a) Changing the Mean



b) Changing the Target Variance



c) Changing the Impostor Variance

Figure A.2: Artificial DET Performance Curves, courtesy of Roland Auckenthaeler

Bibliography

[1] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, Magrin I. Chagnolleau, S. Meignier, T. Merlin, Ortega J. Garcia, Petrovska Delacretaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[2] Yau Wei Yun. *The '123' of Biometric Technology.* http://www.itsc.org.sg/synthesis/2002/biometric.pdf, 2002. as of March 2008.

[3] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and D. A. Reynolds. The nist speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, pages 225–254, 2000.

[4] G. Gravier, J. Kharroubi, and G. Choller. On the Use of Prior Knowledge in Normalisation Schemes for Speaker Verification. In *Digital Signal Processing*, volume 10, pages 213–225, 2000.

[5] J. Thompson. Speech Variability in Speaker Recognition. *PhD Thesis, University of Wales Swansea*, 1997.

[6] Sadaoki Furui. 50 years of progress in speech and speaker recognition. In *Acoustical Society of America Journal*, pages 2497–2498, 2004.

[7] D.A. Reynolds. The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1:113–116 vol. 1, 7-10 May 1996.

[8] J. Ajmera and B. Felix. Age and gender classification using modulation cepstrum. In *In ODYSSEY*, 2008. to be published.

[9] Jonas Richiardi and Andrzej Drygajlo. Evaluation of speech quality measures for the purpose of speaker verification. *In ODYSSEY*, 2008. to be published.

[10] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas. Score normalisation for text-independent speaker verification system. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 Speaker Recognition Workshop*, 10 (1-3):42–54, 2000.

[11] Ganesh N. Ramaswamy Jiri Navrati. *The awe and mystery of T-Norm.* Eurospeech, 2003.

[12] D.E. Sturim and D.A. Reynolds. Speaker adaptive cohort selection for tnorm in text-independent speaker verification. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1:741–744, March 18-23, 2005.

[13] R. Auckenthaler. Text-independent Speaker Verification with Limited Resources. *PhD Thesis, University of Wales Swansea*, 2001.

[14] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alpha nets. In *Proc. ICASSP*, volume 1, pages 397–400, 1991.

[15] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for svm speaker recognition. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1:629–632, March 18-23, 2005.

[16] N. Dehak and G. Chollet. Support vector gmms for speaker verification. *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–4, 28-30 June 2006.

[17] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 1:I–I, 14-19 May 2006.

[18] J. Gauvain and C. Lee. Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains, 1994.

[19] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(2):91–108, August 1995.

[20] A. Zymnis L. Ferrer, M. Graciarena and E. Shriberg. System combination using auxiliary information for speaker verification. *ICASSP*, 2008. to be published.

[21] N. Scheffer and J.-F. Bonastre. Ubm-gmm driven discriminative approach for speaker verification. *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–7, 28-30 June 2006.

[22] R. Stapert. A segmental mixture model: maximising data usage with time sequence information. *PhD Thesis, University of Wales Swansea*, March 2001.

[23] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on ASSP*, 2(4):578–589, 1994.

[24] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, and J. Mason. Influence of task duration in text-independent speaker verification. 2007.

[25] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *Speech and Audio Processing, IEEE Transactions on*, 8(6):695–707, Nov 2000.

[26] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *Speech and Audio Processing, IEEE Transactions on*, 13(3):345–354, May 2005.

[27] K. P. Li and J. E. Porter. Normalizations and selection of speech segments for speaker Recognition Scoring. In *Proc. ICASSP*, pages 595–598, 1988.

[28] D.A. Reynolds. Channel robust speaker verification via feature mapping. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2:II–53–6 vol.2, 6-10 April 2003.

[29] C. Barras and J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2:49–52 vol.2, 6-10 April 2003.

[30] G. Gravier and G. Chollet. Comparison of normalization techniques for speaker verification. *EUROSPEECH*, 1997.

[31] A. E. Rosenberg, J. Delong, C. H. Lee, B. H. Juang, and F. K. Soong. The use of cohort normalised scores for speaker recognition. In *Proc. ICSLP*, pages 599–602, 1992.

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

[33] L.P. Heck and M. Weintraub. Handset-dependent background models for robust text-independent speaker recognition. In *Proc. ICASSP*, 1997.

[34] Ming-Xing Xu Wei Wu, Thomas Fang Zheng and Huan-Jun Bao. *Study on Speaker Verification on Emotional Speech.* In ICSLP, 2006.

[35] Raphaël Blouet Mathieu Ben, Frèdèric Bimbot. A monte-carlo method for score normalization in automatic speaker verification using kullback-leibler distances. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, 2002.

[36] Guillaume Gravier Mathieu Ben, Frèdèric Bimbot. Enhancing the robustness of bayesian methods for text-independent automatic speaker verification. *In ODYSSEY*, 2004.

[37] S. Zhang R. Zheng and B. Xu. A Comparative Study of Feature and Score Normalization for Speaker Verification. *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 531–538, 2005.

[38] Doroteo T. Toledano Cristina Esteve-Elizalde Joaquin Gonzalez-Rodriguez Ruben Fernandez Pozo and Luis Hernandez Gomez. *Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition.* In ODYSSEY, 2008. to be published.

[39] Upendra Chaudhari Jason Pelecanos and Ganesh Ramaswamy. Compensation of utterance length for speaker verification. *In ODYSSEY*, pages 161–164, 2004.

[40] Ignacio Lopez-Moreno Daniel Ramos-Castro, Daniel Garcia-Romero and Joaquin Gonzalez-Rodriquez. *Speaker verification using fast adaptive TNorm based on Kullback-Leibler divergence*, volume 28. Pattern Recognition Letters, ATVS (Speech and Signal Processing Group) Universidad Autonoma de Madrid (Spain), issue 1 edition, 2007.

[41] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece, 1997.

[42] R.B. Dunn, T.F. Quatieri, D.A. Reynolds, and J.P. Campbell. Speaker recognition from coded speech and the effects of score normalization. *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, 2:1562–1567 vol.2, 2001.

[43] D. A. Reynolds. Comparison of Background Normalisation Methods for Text-Independent Speaker Verification. *EuroSpeech97*, 2:963–966, 1997.

[44] W. M. Liu. Objective assessment of comparative intelligibility. *PhD Thesis, Swansea University*, 2008.

[45] J. Koolwaaij, L. Boves, H. Jongebloed, and E. den Os. On model quality and evaluation in speaker verification. *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 6:3759–3762 vol.6, 2000.

[46] J. Gonzalez-Rodriguez D. Garcia-Romero, J. Fierrez-Aguilar and J. Ortega-Garcia. On the use of quality measures for text-independent speaker recognition. *In ODYSSEY*, 2004.

[47] J. Richiardi, P. Prodanov, and A. Drygajlo. A probabilistic measure of modality reliability in speaker verification. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1:709–712, March 18-23, 2005.

[48] Delphine Charlet Mikaël Collet, Yassine Mami and Frédéric Bimbot. *Probabilistic anchor model approach for speaker verification.* Proc. INTERSPEECH, 2005.

[49] NIST. The nist speaker recognition evaluation plan. http://www.nist.gov/speech/tests/sre/, as of March 2008.