Speaker Diarization - "Who Spoke When"

David I-Chung Wang
B.Eng (Hons)

A Thesis Submitted in Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy

at the

Queensland University of Technology
Speech and Audio Research Laboratory
Science and Engineering Faculty

October 2012

Keywords

Speaker diarization, speaker segmentation, speaker clustering, Bayes factors, eigenvoice modelling, joint factor analysis, distance measures.

Abstract

Speaker diarization is the process of annotating an input audio with information that attributes temporal regions of the audio signal to their respective sources, which may include both speech and non-speech events. For speech regions, the diarization system also specifies the locations of speaker boundaries and assign relative speaker labels to each homogeneous segment of speech. In short, speaker diarization systems effectively answer the question of 'who spoke when'.

There are several important applications for speaker diarization technology, such as facilitating speaker indexing systems to allow users to directly access the relevant segments of interest within a given audio, and assisting with other downstream processes such as summarizing and parsing. When combined with automatic speech recognition (ASR) systems, the metadata extracted from a speaker diarization system can provide complementary information for ASR transcripts including the location of speaker turns and relative speaker segment labels, making the transcripts more readable. Speaker diarization output can also be used to localize the instances of specific speakers to pool data for model adaptation, which in turn boosts transcription accuracies. Speaker diarization therefore plays an important role as a preliminary step in automatic transcription of audio data.

The aim of this work is to improve the usefulness and practicality of speaker diarization technology, through the reduction of diarization error rates. In particular, this research is focused on the segmentation and clustering stages within a diarization system. Although particular emphasis is placed on the broadcast news audio domain and systems developed throughout this work are also trained and tested on broadcast news data, the techniques proposed in this dissertation are

also applicable to other domains including telephone conversations and meetings audio.

Three main research themes were pursued: heuristic rules for speaker segmentation, modelling uncertainty in speaker model estimates, and modelling uncertainty in eigenvoice speaker modelling.

The use of heuristic approaches for the speaker segmentation task was first investigated, with emphasis placed on minimizing missed boundary detections. A set of heuristic rules was proposed, to govern the detection and heuristic selection of candidate speaker segment boundaries. A second pass, using the same heuristic algorithm with a smaller window, was also proposed with the aim of improving detection of boundaries around short speaker segments. Compared to single threshold based methods, the proposed heuristic approach was shown to provide improved segmentation performance, leading to a reduction in the overall diarization error rate.

Methods to model the uncertainty in speaker model estimates were developed, to address the difficulties associated with making segmentation and clustering decisions with limited data in the speaker segments. The Bayes factor, derived specifically for multivariate Gaussian speaker modelling, was introduced to account for the uncertainty of the speaker model estimates. The use of the Bayes factor also enabled the incorporation of prior information regarding the audio to aid segmentation and clustering decisions.

The idea of modelling uncertainty in speaker model estimates was also extended to the eigenvoice speaker modelling framework for the speaker clustering task. Building on the application of Bayesian approaches to the speaker diarization problem, the proposed approach takes into account the uncertainty associated with the explicit estimation of the speaker factors. The proposed decision criteria, based on Bayesian theory, was shown to generally outperform their non-Bayesian counterparts.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
CLR	Cross Likelihood Ratio
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DER	Diarization Error Rate
DET	Detection Error Trade-off
EARS	Effective, Affordable, Reusable Speech-to-text
E-M	Expectation-Maximisation algorithm
FFT	Fast Fourier Transform
GLR	Generalized Likelihood Ratio
GMM	Gaussian Mixture Model
iid	independent and identically distributed
INA	Institut National de l'Audiovisuel

LIST OF ACRONYMS AND ABBREVIATIONS

JFA	Joint Factor Analysis
KL	Kullback-Leibler divergence
LDC	Linguistic Data Consortium
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LP	Linear Predictor
LPCC	Linear Predictive Cepstral Coefficients
MAP	Maximum *A Posteriori*
MDE	Metadata Extraction
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
NIST	National Institute of Standards and Technology
PLP	Perceptual Linear Predictive coefficients
RT	Rich Transcription
SAD	Speech Activity Detection
SRE	NIST Speaker Recognition Evaluation
UBM	Universal Background Model

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: _____

Date: _____

Acknowledgements

First and foremost, I would like to thank my immediate family, in particular my lovely parents and my wonderful partner Lee. Your unconditional love and support means the world to me. Thank you for believing in me, and for your constant encouragement along the way. Thank you for allowing me to live worry-free, and for putting up with me through the stressful times, when I'm probably less than pleasant to be around. To my partner and best friend Lee, thank you for giving me something to look forward to at the end of each day, and thank you for the fun times, the memorable trips, and the discussions about our future goals. I couldn't have done it without you by my side.

I would like to thank my principal supervisor, Professor Sridha Sridharan, for giving me the opportunity to undertake this research programme, and for providing equipment, IT support and research funding. Your encouragement and support throughout this work is also most appreciated.

A special thanks is reserved for Dr Robert Vogt, who I'm very fortunate to have as my associate supervisor. Thank you for showing me the ropes and providing technical guidance every step of the way. Thank you for answering my many and sometimes not-so-intelligent questions. Without your support, ideas, technical expertise and *infinite* patience, this work would not have been possible.

To all my friends in and outside the SAIVT lab, in particular Brenden, Daniel, Ellan and Brian - thank you for keeping me sane, and giving me a much needed break in the form of social life, especially during the write-up of this thesis. And thanks for *not* asking how my PhD is going on a regular basis!

Chapter 1

Introduction

1.1 Motivation and Overview

With the ever increasing number of TV channels and broadcasting radio stations and the continually decreasing cost and increasing volume of large storage means, large volumes of spoken documents, such as broadcast news, are being recorded and audio archives around the world are expanding on a daily basis. For example, the *Institut National de l'Audiovisuel* (INA) in France possesses 45 years of TV archives consisting of 300,000 hours of national TV programs, and 60 years of radio archives consisting of 400,000 hours of radio programs [15]. Compared to text, accessing and searching audio documents is a much more difficult and time consuming task. There is therefore a growing need to apply automatic human language technologies to allow efficient searching, indexing and accessing of these information sources.

One obvious solution to this task is the use of speech-to-text technology to transcribe the audio data. Indeed, automatic speech recognition (ASR) is an active field of research that have received significant attention in the speech processing community for decades. ASR systems provide a sound base for these tasks; however the transcripts are often difficult to read, and do not capture all the information contained within the audio, such as the location of speaker turns and the speaker segment labels. In addition to the fundamental speech recogni-

tion technologies, other technologies are therefore required to extract these *metadata* to provide information beyond the words that were spoken. The availability of metadata offers the potential to provide more context and thus making the transcripts more readable, as well as assist with other downstream tasks such as summarizing and parsing.

Apart from the aforementioned advantages, metadata regarding the location of speaker turns and the speaker segment labels can also be used for information retrieval purposes, allowing the user to formulate a query in a convenient way and to find the searched information from the audio recording quickly and effectively. For example, a user may wish to locate instances of a specific speaker within a broadcast news program. It is impractical and time consuming for the user to listen to the whole audio, which may contain music segments, commercials and speech from irrelevant speakers, in order to retrieve pertinent information. In this case, information regarding the location of speaker turns and the speaker segment labels can be used to facilitate a speaker indexing system, which allows the user to directly access the relevant segments of interest, as well as provide functionality such as 'how many speakers are in the audio', 'when did this person speak', 'how long did this person speak for' and 'show me what else this person said'.

The process of detecting speaker turns and providing relative labels for each speaker segment is known as speaker segmentation and clustering, or collectively as speaker diarization, and is the topic of this dissertation.

1.1.1 Speaker Diarization

More formally, a speaker diarization system aims to annotate an input audio with information that attributes temporal regions of the audio signal to their respective sources, which may include both speech and non-speech events. For speech regions, the diarization system also specifies the locations of speaker boundaries and assign relative speaker labels to each homogeneous segment of speech. In short, speaker diarization systems effectively answer the question of 'who spoke

when'.

In recent years, speaker diarization has emerged as an increasingly important and dedicated domain of speech research. Apart from facilitating the searching and indexing of audio archives and making ASR transcripts more readable, speaker diarization systems have also proven useful as a preprocessing step to assist with the speech recognition process itself. In generic ASR systems, universal speech unit models used for recognition is generally trained on large databases with a large number of speakers. In this case, enhanced recognition performance can be obtained by using speaker labels produced by a diarization system to localize the instances of specific speakers to pool data for model adaptation, which in turn boosts transcription accuracies. Moreover, diarization systems can also locate non-speech events such as music, which can then be ignored by the ASR system to reduce computational costs. Speaker diarization hence plays an important role as a preliminary step in automatic transcription of audio data.

Despite a rapid increase in popularity over recent years, the field of speaker diarization research is still relatively unexplored compared to the mature fields of speech processing research, such as speech recognition and speaker verification.

1.2 Aim and Scope

Given the practical applications of speaker diarization systems, the overall aim of this work is to improve the usefulness of speaker diarization technology by investigating novel approaches with the goal of reducing diarization error rates.

There are three primary domains which have been used for speaker diarization research and development over the years. These are broadcast news audio, telephone conversations and recorded meetings. Each of these application domains presents unique diarization challenges, through the different number of speakers present in the audio, the quality of the recordings, the amount and types of background noise, the duration and sequencing of speaker segments, the amount of overlapping speech that is likely to occur, and whether the speech is scripted

or spontaneous. The work presented in this research will focus on the broadcast news domain, and systems developed throughout this work will be trained and tested only on broadcast news data. Despite this, however, it is worth noting that the techniques proposed in this dissertation are also applicable to other audio domains.

As the broadcast news shows are generally well scripted and overlapping speech rarely occurs, the detection of overlapping speech will not be investigated in this research.

Within the broadcast news domain, the scope of this work will be restricted to the underlying technology of speaker diarization, that is, the pattern recognition techniques associated with locating speaker change points within the audio, and associating speech segments belonging to the same speaker. Although speech activity detection is commonly performed as a preprocessing step in a typical speaker diarization system, it is considered a separate topic and is outside the scope of this dissertation.

Throughout this work, it will be assumed that no prior knowledge is available regarding the input audio, including the number of speakers, the identity of the speakers and the structure of the news shows. For each show, the speaker diarization system therefore outputs relative speaker labels for each homogeneous speaker segment, such as 'speaker1' or 'speaker2', rather than absolute speaker labels that associate each speaker segment with the identity of the speaker. This requirement is consistent with the protocols defined in the speaker diarization metadata extraction (MDE) task of the U.S. National Institute of Standards and Technology (NIST) Rich Transcription (RT) Evaluations [19]. Incorporating prior information regarding the speaker identities and having prior access to samples of speech from one or more speakers within the audio is associated with speaker detection or tracking tasks, which fall outside the scope of this work.

The practicalities of deploying a speaker diarization system are also beyond the scope of this research, as is the integration of the diarization system with other existing systems and downstream processes that can benefit from diarization

output. The development of a speaker diarization user interface will not be included as a part of this research programme.

To achieve the aim of this research programme, three avenues for the development and improvement of speaker diarization technologies are pursued.

Heuristic rules for speaker segmentation: A popular approach for speaker segmentation involves the use of distance metrics to determine how statistically similar the audio is on either side of a given location within an audio, to determine how much a segment boundary is favoured at that particular point. Segment boundaries are then hypothesized based on the location of local maxima of the resultant distance curve. Due to the nature of speech signals, the resultant distance curve is intrinsically noisy, and not all local maxima of the distance curve correspond to true speaker change points. The use of a simple threshold for boundary decisions is thus insufficient to ensure reliable segmentation. Instead, appropriate heuristic rules can be applied to determine which local maxima correspond to true speaker change points. Heuristic rules for metric based segmentation have not been investigated in detail in speaker diarization literature.

Modelling uncertainty in speaker model estimates: For speaker segmentation and clustering, speaker models are first estimated using data contained in the speaker segments of interest, and distance metrics are used to make segmentation and clustering decisions. The distance metrics proposed in literature do not take into account the uncertainty associated with the speaker model estimates. Theoretically, this would not be an issue if an infinite amount of speech is present in the speaker segments, given that the model parameters estimated from an infinite amount of data would be very close to the true parameters. However, in practice, the data in each speaker segment is always limited, and there is always a degree of uncertainty associated with the estimation of each speaker model. This uncertainty therefore needs to be taken into consideration, especially for short segments which contain less data, thus resulting in a higher degree of uncertainty.

Moreover, the use of prior information regarding the audio to aid segmentation and clustering is not well explored in literature. Although the scope of this work assumes that no prior information regarding the audio is available, the diarization process can still retrieve 'prior' information from the given audio data itself, to determine what a generic speaker model in that particular audio should look like. This information can then be used to aid the segmentation and clustering tasks.

Modelling uncertainty in eigenvoice speaker modelling: Recently, the eigenvoice modelling framework was introduced as an alternative to traditional Gaussian mixture model (GMM) based speaker modelling. One major advantage of this approach is the ability to estimate more reliable speaker models using limited enrolment data. Once again, the uncertainty associated with the explicit estimation of the speaker factors is not taken into account in this approach. The concept of modelling the uncertainty of speaker model estimates explicitly can therefore be extended to this modelling framework.

1.3 Thesis Structure

The remaining chapters of this thesis are composed as follows:

Chapter 2 provides an overview of existing speaker diarization research, as well as common speaker modelling techniques reported in literature. Significant attention is given to the speaker segmentation and clustering tasks within a diarization system, which is the focus of the research presented in this dissertation.

Chapter 3 discusses the evaluation metrics used throughout this dissertation to measure and compare the performance of developed diarization systems. Various speech corpora used for training speaker models and testing the performance of developed systems are also described, and evaluation pro-

tocols are defined. Finally, a baseline broadcast news diarization system is described. This system will be used for performance comparisons throughout this work.

Chapter 4 outlines some preliminary investigations into the use of heuristic rules for speaker segmentation, with emphasis placed on minimizing missed boundary detections. The proposed heuristic rules determine which local maxima points (or peaks) on the distance curve correspond to true speaker change points, based on analysing the relative depths of the local minimum (or trough) between successive peaks, as well as the absolute value of the distance score at the location of the peak itself. The incorporation of a smaller analysis window is also investigated, with the aim of improving boundary detection between short speaker segments.

Chapter 5 develops the Bayes factor as a decision criterion for speaker segmentation and clustering, as an alternative to the popular maximum likelihood based metrics reported in literature. Rather than treating the estimated model parameters as known variables, the Bayes factor approach assumes that the model parameters are random variables, each with their own probability distributions. The Bayes factor approach is able to model the uncertainty associated with the estimation of model parameters, as well as incorporate prior knowledge about the audio to aid segmentation and clustering decisions.

Chapter 6 extends the concept of modelling uncertainty in speaker model estimates to the eigenvoice modelling framework for speaker clustering. Building on the success of eigenvoice modelling of speaker segments for diarization, the proposed approach is able to model the uncertainty associated with the explicit estimation of speaker factors, effectively combining the eigenvoice modelling framework with the advantages of a Bayesian approach to the diarization problem.

Chapter 7 concludes this dissertation with a summary of the contributions of

this research programme and suggests further directions for continuing research in the field of speaker diarization.

1.4 Original Contributions of this Thesis

This research programme has resulted in contributions to the field of speaker diarization in all of of the research themes identified above.

1.4.1 Heuristic Rules for Speaker Segmentation

Heuristic rules for metric based segmentation were developed, to determine which peaks on the distance curve correspond to true speaker change points. The proposed heuristic rules govern the smoothing of the distance curve, detection of peaks on the smoothed curve and the selection of significant peaks as segment boundary hypotheses. A novel method is proposed for selecting significant peaks, based on analysing the depth of the valleys between peaks.

To minimize missed boundary detections between short speaker segments, a second pass is proposed using a smaller analysis window. A minimum segment duration constraint is imposed to ensure that the boundaries detected in the first pass are not detected again in the second pass as a different boundary located in close proximity to the initial boundary.

Preliminary experiments demonstrate that the proposed heuristic algorithm achieves improved segmentation performance compared to the baseline system which uses a simple threshold for segment boundary decisions. This leads to an improvement in the overall diarization performance.

1.4.2 Modelling Uncertainty in Speaker Model Estimates

To account for the uncertainty associated with the direct estimation of model parameters, the Bayes factor is proposed as a decision criterion for speaker segmentation and clustering, replacing the popular maximum likelihood criteria widely

reported in literature. The Bayes factor approach is able to incorporate the information regarding the whole audio as prior information, to aid segmentation and clustering decisions.

Considering the speaker clustering task in a hypothesis testing framework, a solution to the Bayes predictive density for single, full covariance multivariate Gaussian speaker models is derived, which forms the basic building block for constructing the Bayes factor criterion. The derivations of the solution to the Bayes predictive density integral invokes a technique known as simultaneous diagonalization, to transform the data such that the covariance matrices of the data and prior distributions are simultaneously diagonal in the transformed space. A method of estimating the hyperparameters is also proposed, allowing the knowledge regarding the audio to be effectively incorporated as prior information to aid clustering decisions. The Bayes factor can be used to replace the popular Bayesian information criterion (BIC) commonly reported in literature for speaker clustering tasks. A derivation of the BIC is also presented, showing the BIC as only an approximation to the Bayes factor itself.

The concept of using the Bayes factor for speaker clustering is then extended to the segmentation task. The solution to the Bayes predictive density integral is used to construct a Bayes factor as a distance metric that is more suitable for this application, replacing the popular generalized likelihood ratio (GLR).

1.4.3 Modelling Uncertainty in Eigenvoice Speaker Modelling

An alternative approach to speaker modelling, based on joint factor analysis (JFA) techniques [40], is the use of the eigenvoice model to represent the speakers [41, 44, 62]. Under this approach, the GMM that best represents the observations of a particular speaker is given by the combination of a speaker-independent universal background model (UBM) with an additional speaker-dependent offset constrained to lie in a low-dimensional speaker variability subspace.

A novel criterion is proposed in this work, by incorporating eigenvoice mod-

elling techniques into the popular cross likelihood ratio (CLR) criterion [5, 71] for speaker clustering, allowing the system to capitalize on the advantages of eigenvoice modelling techniques in a CLR framework. The proposed eigenvoice-CLR approach replaces the traditional CLR based clustering using GMM modelling of speaker segments.

Building on the idea of employing Bayesian approaches for the speaker clustering task, which allows the uncertainties of speaker model estimates to be taken into account, a Bayesian version of the eigenvoice-CLR criterion is also developed in this work. This novel criterion uses Bayesian methods to estimate the conditional probabilities in computing the CLR, effectively combining the eigenvoice-CLR framework with the advantages of a Bayesian approach to the diarization problem.

In order to incorporate knowledge of the audio into the clustering decision effectively, it is proposed to transform the supervector space to centre the origin at the location of the background mean supervector. It is discovered that the proposed Bayesian eigenvoice-CLR criterion essentially reverts back to the form of a log likelihood ratio in the transformed space, with the exception of a multiplicative normalization constant. As the application of the normalization constant in the CLR criterion appears to be unintuitive, an investigation is conducted to observe the effects of removing this normalization constant on the performance of the clustering system. This is followed by an extensive analysis of the importance of the normalization constant.

Experimental results demonstrate the robustness of the proposed Bayesian eigenvoice-CLR criterion, which significantly improves diarization performance over the baseline system. Intermediate systems are also developed to indicate how much of the overall improvement can be attributed to each successive refinement stage leading up to the final Bayesian eigenvoice-CLR criterion. These include the replacement of traditional MAP adaptation of Gaussian mixture speaker models with eigenvoice adaptation, the integration of eigenvoice modelling techniques into the CLR framework, and employing a Bayesian approach for evaluating the

conditional probabilities in the eigenvoice-CLR criterion.

The incorporation of a residue term into the speaker model provides additional modelling power through the introduction of extra model parameters. The residue term aims to model any residual speaker variations that the speaker factor term fails to take into account. It is discovered that further improvements to diarization performance can be achieved by incorporating the residue term.

1.5 Publications

Listed below are the publications resulting from this research programme.

Peer-reviewed International Journal

- **D. Wang**, R. Vogt and S. Sridharan, "Eigenvoice modeling for cross likelihood ratio based speaker clustering: a Bayesian approach," *Computer Speech and Language*, 2012. (Submitted)

Peer-reviewed International Conferences

- **D. Wang**, R. Vogt, S. Sridharan and D. Dean, "Cross likelihood ratio based speaker clustering using eigenvoice models," in *Interspeech*, 2011, pp. 957-960.

- **D. Wang**, R. Vogt and S. Sridharan, "Bayes factor based speaker segmentation for speaker diarization," in *Interspeech*, 2010, pp. 1405-1408.

- **D. Wang**, R. Vogt and S. Sridharan, "Bayes factor based speaker clustering for speaker diarization," in *International Conference on Information Sciences, Signal Processing and their Applications (ISSPA)*, 2010, pp. 61-64.

- **D. Wang**, R. Vogt, M. Mason and S. Sridharan, "Automatic audio segmentation using the generalized likelihood ratio," in *International Conference*

on Signal Processing and Communication Systems (ICSPCS), 2008, pp. 1-5.

Chapter 2

An Overview of Speaker Diarization Technology

2.1 Introduction

Audio diarization is the process of annotating an input audio stream with information that attributes temporal regions of the audio signal into their specific sources. These sources can include speakers, background noise, music, environmental and channel characteristics [74].

In its simplest form, an audio diarization system breaks the input audio into speech and non-speech segments, which include silence, music and noise. A more complicated diarization system would also indicate the location of speaker changes within speech segments, and associate segments of speech coming from the same speaker. This is referred to as speaker diarization, and is the focus of most recent research efforts in audio diarization and is also the topic of this dissertation.

The U.S. National Institute of Standards and Technology (NIST) has sponsored a number of internationally competitive evaluations which have contributed enormously to the advancement of speaker diarization technology. These official benchmark evaluations are an important vehicle for driving the state-of-the-art diarization technology forward, by defining standardized experimental protocols and databases, thus allowing meaningful comparisons to be made between var-

ious proposed techniques. Over the years, the speech processing research community has developed speaker diarization systems in a number of different audio domains. Prior to 2002, speaker diarization was evaluated through the NIST Speaker Recognition Evaluations (SRE), which focused on conversational telephone speech data. From 2002 onwards, the speaker diarization task was evaluated as a part of the Rich Transcription (RT) Evaluations, which primarily focused on broadcast news data between 2002 and 2004. Since 2005, the meetings audio domain has received the most attention. Each of these application domains presents unique diarization challenges, through the different number of speakers present in the audio, the quality of the recordings, the amount and types of background noise, the duration and sequencing of speaker segments, the amount of overlapping speech that is likely to occur, and whether the speech is scripted or spontaneous. An extensive analysis of the characteristics of broadcast news audio is presented in [50]. A comparison of broadcast news and meetings audio can be found in [3].

The work presented in this dissertation will focus on the broadcast news audio domain. Despite the fact that the most recent NIST RT Evaluations have primarily focused on meetings audio, an ideal solution to the speaker diarization problem for broadcast news is just as valuable, and is still far from being established. Moreover, many techniques proposed in speaker diarization literature tend to generalize well over different audio domains [74]. As stated in Section 1.2, although this work will focus on broadcast news data, the techniques proposed throughout this dissertation are also applicable to other audio domains.

This chapter provides an introduction to the speaker diarization problem and serves as a review of existing techniques reported in speaker diarization literature. Section 2.2 outlines a generic, prototypical speaker diarization system, which can be broken down into three stages, namely speech activity detection (SAD), speaker segmentation and speaker clustering. Before any techniques reported in literature for any of those stages can be described in more detail, it is useful to first outline the feature extraction task as well as the speaker modelling techniques

commonly used in speaker diarization systems today. Section 2.3 outlines some commonly used features that are extracted from the raw speech signal to obtain the speaker specific information. Section 2.4 reviews the theory behind Gaussian mixture speaker modelling, which is the dominant modelling process for speaker recognition applications. Approaches to estimating the Gaussian mixture model parameters are also detailed. While still based on the idea of using Gaussian mixture models (GMM) to model a speaker, recent advances in speaker modelling through the use of eigenvoice modelling techniques is described in Section 2.5.

Following the description of the feature extraction task and speaker modelling approaches, the discussion returns to the three main stages within a speaker diarization system in Sections 2.6 to 2.8, detailing the popular existing techniques reported in literature for each stage. For the speaker segmentation and clustering stages, much of the discussion is centred around the use of distance metrics in making segmentation and clustering decisions. The distance metrics measure the degree of statistical dissimilarity between two segments or clusters of interest, and is the focus of much of this dissertation. Finally, eigenvoice modelling based speaker diarization systems reported in recent literature is described in Section 2.9.

2.2 Automated Speaker Diarization

The basic speaker diarization process generally comprises of three main components, namely speech activity detection, speaker segmentation and speaker clustering, as shown in Figure 2.1. While some modern approaches to speaker diarization tackle the segmentation and clustering tasks simultaneously, both of these modules are fundamental to the speaker diarization problem, and some recent systems such as [85] still apply distinctly independent segmentation and clustering stages. This work will thus consider speaker segmentation and clustering as separate components. The role of each component shown in Figure 2.1 is as follows:

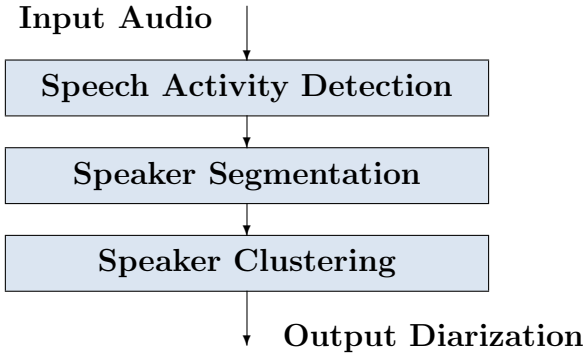

Figure 2.1: The basic speaker diarization process

Speech Activity Detection: The purpose of this stage is to classify the audio into speech and non-speech regions. It is important to identify and discard non-speech regions such as music and noise early in the diarization process, to avoid hindering subsequent speaker segmentation and clustering processes. The aim of this stage is to remove only prolonged periods of music or noise, rather than targeting short speaker pauses in the middle of speaker turns and thus breaking up homogeneous speaker segments. All audio regions classified as non-speech are excluded from further processing.

Speaker Segmentation: Speaker segmentation is the process of partitioning the audio data into homogeneous segments according to speaker identities. This stage is responsible for determining all boundary locations within each speech region that correspond to true speaker change points, providing clean, uncontaminated data for subsequent speaker clustering.

Speaker Clustering: Speaker clustering is the process of associating segments of speech produced by the same speaker. This stage is responsible for labelling all speech segments belonging to the same speaker with the same relative, show-internal speaker label. Ideally, at the end of this stage, one single cluster is produced for each speaker in the audio, containing all speech segments belonging to that speaker.

Existing techniques reported in speaker diarization literature for each of these stages generally rely on the use of suitable speaker characterising features to convey speaker specific information, as well as the appropriate modelling of the

distribution of the extracted features to construct models that define the 'voice print' of each speaker. These topics will be discussed in Sections 2.3 to 2.5.

2.3 Feature Extraction

In speaker recognition applications, speech processing is not commonly performed directly on the raw audio signal. Most applications, including all speaker segmentation and clustering techniques covered throughout this dissertation, require features containing speaker-specific information to be extracted from the audio signal, to allow subsequent speaker modelling and classification to be performed using these features. The extracted features should ideally maximise inter-speaker variability while minimizing intra-speaker variations, and represent the relevant information in a compact form [18].

To date, cepstral-domain features based on short periods of speech have proven to be most successful at capturing the useful characteristics of speech for speech and speaker recognition applications, compared to both time-domain signals and frequency-domain spectra [57]. Cepstral features are capable of capturing information pertinent to the anatomical aspects of the speaker's voice production mechanisms, including the vocal tract shape and glottal source. This allows acoustical and physical traits such as their tone, the nasality and the roughness of their voice to be encoded in the features. This class of features include mel-frequency cepstral coefficients (MFCC) [14], linear predictive cepstral coefficients (LPCC) [23], and perceptual linear predictive (PLP) coefficients [37]. A detailed evaluation of these acoustic features for speaker recognition can be found in [58].

Cepstral features are based on spectral information from relatively short segments of speech, referred to as frames, which typically contain 10-30 milliseconds of speech with a significant overlap between consecutive frames. It is assumed that the speech signals are quasi-stationary within the short periods considered. A sliding spectral analysis window [7] (not to be confused with the sliding-window method for speaker segmentation described in Section 2.7) is applied to the frames

of speech to provide a more consistent response across all frequencies and pitches of speech. A Hamming window is used in this work as is typically employed in literature. The windowed frames of speech are then used to compute a sequence of magnitude spectra, and the spectral representations are then transformed to cepstral coefficients as a final step. Each frame of speech results in a single feature vector.

Two categories of cepstral features are commonly used in speaker recognition systems reported in literature, differing in the method by which the log-magnitude spectrum is represented. Filterbank analysis describes the magnitude spectrum through the energy in the output signal of a set of bandpass filters, while linear predictive analysis involves approximating the magnitude spectrum using an all-pole filter.

2.3.1 Filterbank Analysis

Although filterbank analysis was one of the earliest methods developed for speech processing, it remains one of the most effective techniques used in speaker recognition systems today [46]. In this approach, the short-time magnitude spectrum of a speech signal is represented by the energy in the output signal of a set of bandpass filters spaced evenly across the frequency range of interest. Approximately 20 filters are typically used in this process, producing a compact set of coefficients to represent the spectrum.

Based on the filterbank analysis approach, mel-frequency cepstral coefficients (MFCC) are the most popular and commonly used of the acoustic features and it has been demonstrated to work well in both speech recognition [14] and speaker recognition [46] tasks. MFCCs are produced through spacing the filters evenly according to the mel-frequency scale. The mel-frequency scale is a non-linear transformation of the physical frequency to the pitch perceived by humans [72], placing less emphasis on higher frequencies. In this way, the bandwidth of each filter represents a perceptually similar frequency range and quantity of information content. The mel-frequency scale is logarithmic in the standard frequency

scale and is approximated by

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right). \tag{2.1}$$

For computational efficiency, the filterbank is implemented in the frequency domain using the fast Fourier transform (FFT) of the speech frames.

The log-energies of the filterbank outputs are transformed into cepstral coefficients via the discrete cosine transform (DCT). This significantly reduces the correlation between the energy outputs of adjacent (and usually partially overlapping) bandpass filters, thus allowing simpler subsequent modelling of speech using these feature vectors.

The time derivatives of the static features, also known as delta coefficients, are often appended to the feature vector as additional features to model trajectory information [10]. Delta coefficients approximate the instantaneous derivative of each of the cepstral coefficients by finding the slope coefficient via least-square linear regression over a window of consecutive frames. The window lengths are typically between 3 and 7 frames.

2.3.2 Linear Predictive Analysis

In linear predictive analysis, the speech production model is assumed to incorporate a glottal excitation signal filtered through the vocal tract and nasal cavity. Let the speech signal at time n be denoted by s_n. The linear predictor (LP) models s_n by a linear combination of its past values and a weighted present input excitation [45] as

$$s_n = G \cdot u_n - \sum_{k=1}^{P} a_k \cdot s_{n-k}\,, \tag{2.2}$$

where G is a gain scaling factor, u_n is the present input excitation, P is the prediction order, and a_k is a set of model parameters called the predictor coefficients, which define an all-pole filter that describes the response of the vocal tract given an input excitation signal. The predictor coefficients are estimated

using a minimum mean squared error (MMSE) criterion where the residual error is assumed to be equivalent to the excitation term, Gu_n. While some speaker-dependent characteristics such as the fundamental frequency of voiced speech can be extracted from this excitation term [10], it is the predictor coefficients that are usually the part of the model of interest in feature extraction, as they provide the majority of speaker discriminative information.

In order to express the features in a more appropriate form for speaker modelling, the predictor coefficients calculated based on linear predictive analysis are commonly transformed into linear predictive cepstral coefficients (LPCC) [23], which have found significant use in speaker recognition tasks [5, 61]. Similar to MFCCs, LPCC features are derived through a further Fourier or cosine transform from the log-magnitude of the spectrum, and delta coefficients are commonly appended to the feature vector to capture transient information. However, the log-magnitude of the spectrum in this case is estimated via the frequency response of the all-pole filter defined by the predictor coefficients.

Based on LP modelling, the perceptual linear predictive (PLP) analysis technique [37] attempts to represent speech based on human perception by incorporating several human perceptual factors to the speech signal before applying the LP model. Similar to the mel-frequency transformation for calculating MFCCs, a Bark-scale transformation is applied to the power spectrum in order to equalize the information content of the signal. Additionally, the difference in perceived loudness for different frequencies and power levels are also normalized.

2.4 Gaussian Mixture Speaker Modelling

Once the appropriate features have been extracted from a speaker's speech, GMMs can be trained to represent the speaker by modelling the distribution of the feature vectors. A GMM is a parametric model that can be used to estimate continuous probability density functions for multi-dimensional features. It is widely accepted today that GMMs are one of the more successful structures

used for modelling the statistical characteristics of a speaker. The use of GMMs for speaker modelling have been reported in speaker processing literature for a wide range of applications, including speaker diarization [49, 53, 83], speaker identification [64] and speaker verification [60] tasks.

Given a single feature vector observation $\boldsymbol{x}$, the probability density of $\boldsymbol{x}$ from a D-dimensional multivariate Gaussian distribution is given by

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{2.3}$$

where $\boldsymbol{\mu}$ is the distribution mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix.

In order to model the complex distributions of feature vectors in more detail, a probabilistic model known as mixture distributions [47] can be formulated, where the overall probability distribution is given by an additive linear combination of multiple weighted component densities. A mixture distribution with Gaussian densities as basis components is therefore referred to as the Gaussian mixture model. By adjusting the means, covariances and weights for each Gaussian component, almost any continuous, arbitrarily shaped densities can be approximated to a desired accuracy provided sufficient components are used [8], making the GMM a very powerful modelling structure.

For a GMM with C component densities, the speaker model can be described in full by specifying the set of parameters λ, which consist of mixture component weights ω_c, mean vectors $\boldsymbol{\mu}_c$ and covariance matrices $\boldsymbol{\Sigma}_c$, that is, $\lambda = \{\omega_1, \ldots, \omega_C, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_C\}$. The mixture weights satisfy the constraint

$$\sum_{c=1}^{C} \omega_c = 1. \tag{2.4}$$

For the single feature vector observation $\boldsymbol{x}$, the overall probability density given a GMM described by model λ is given by

$$p(\boldsymbol{x}|\lambda) = \sum_{c=1}^{C} \omega_c g(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c). \tag{2.5}$$

For speaker recognition tasks, the covariance matrices $\boldsymbol{\Sigma}_c$ can be fully specified, or alternatively a diagonal constraint may be imposed. The choice between full or diagonal covariance matrices as well as the number of Gaussian components depends on the application and the amount of data available for estimating the GMM parameters. It is worth noting that since the component Gaussian densities are combined together to model the overall feature density, full covariance matrices are not essential even if the features are not statistically independent. The linear combination of diagonal covariance basis Gaussian densities is capable of modelling the correlations between feature vector elements. The density modelling of a full covariance mixture can therefore be equally well achieved using a higher order, diagonal covariance mixture [59].

For an entire utterance, each feature vector observation is assumed to be independent and identically distributed. The density for an utterance $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ of length T is therefore given by the joint density

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda) = \prod_{t=1}^{T} \sum_{c=1}^{C} \omega_c g(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\,. \tag{2.6}$$

The following sections describe the approaches for estimating the parameters of a GMM, based on maximum likelihood and maximum *a posteriori* criteria.

2.4.1 Maximum Likelihood Estimation

The aim of maximum likelihood (ML) estimation is to find the set of model parameters which maximize the likelihood of the training data. For single multivariate Gaussian modelling, the estimation of model parameters using the maximum likelihood criterion is a relatively trivial task. However, for GMMs, this task is not so straightforward due to the assumptions made for mixture models. Central to the mixture model concept is the assumption that any given observation was produced by only one component of the mixture [47]. The resulting issue in the estimation procedure is that the Gaussian mixture component responsible for producing each feature vector is unknown.

The expectation-maximisation (E-M) algorithm [16] is a well-known technique designed for this situation, where the data is missing or incomplete. The E-M algorithm is an iterative procedure which aims to estimate a new and improved set of model parameters λ using the training data $\boldsymbol{X}$ and the current model parameters $\hat{\lambda}$. In each iteration, the aim is to provide a new and improved estimate of the model, such that $p(\boldsymbol{X}|\lambda) \geq p(\boldsymbol{X}|\hat{\lambda})$ [64]. As the name suggests, the algorithm consists of two steps, the expectation step (or E-step) and the maximisation step (or M-step). The E-step attempts to complete the data by evaluating the *expected* value of the missing data based on the current estimate of the model parameters and the training data. The M-step is then responsible for adjusting the current model parameter estimates using this information.

Specifically, in the case of a GMM, the missing information is the mixture component labels that produced each observation. The E-step therefore involves calculating the expected probability of the observation $\boldsymbol{x}$ being produced by mixture component c using the current model estimate $\hat{\lambda}$, via

$$P(c|\boldsymbol{x}) = \frac{\hat{\omega}_c g(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)}{p(\boldsymbol{x}|\hat{\lambda})} \,. \tag{2.7}$$

Next, the M-step aims to use this information to refine the current model parameters, through maximising the auxiliary function $Q(\lambda; \hat{\lambda})$, given by

$$Q(\lambda; \hat{\lambda}) = \sum_{t=1}^{T} \log \left(\sum_{c=1}^{C} P(c|\boldsymbol{x}_t) \omega_c g(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \right) . \tag{2.8}$$

The direct maximisation of this expression is difficult due to the log of a sum. A simpler auxiliary function $\tilde{Q}(\lambda; \hat{\lambda})$ can be produced by invoking Jensen's inequality [36], which ensures that $Q(\lambda; \hat{\lambda}) \geq \tilde{Q}(\lambda; \hat{\lambda})$. It can be shown [8] that maximising $\tilde{Q}(\lambda; \hat{\lambda})$ for the model parameters λ guarantees that $p(\boldsymbol{X}|\lambda) \geq p(\boldsymbol{X}|\hat{\lambda})$. The simpler auxiliary function is given by

$$\tilde{Q}(\lambda; \hat{\lambda}) = \sum_{t=1}^{T} \sum_{c=1}^{C} P(c|\boldsymbol{x}_t) \log \omega_c g(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \,. \tag{2.9}$$

Maximising (2.9) for the GMM parameters results in the following update equations

$$\omega_c = \frac{n_c}{T} \tag{2.10}$$

$$\boldsymbol{\mu}_c = \frac{1}{n_c}\boldsymbol{S}_{X;c} \tag{2.11}$$

$$\boldsymbol{\Sigma}_c = \frac{1}{n_c}\boldsymbol{S}_{XX;c} - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^{\mathrm{T}}\,, \tag{2.12}$$

using the statistics

$$n_c = \sum_{t=1}^{T} P(c|\boldsymbol{x}_t, \hat{\lambda}) \tag{2.13}$$

$$\boldsymbol{S}_{X;c} = \sum_{t=1}^{T} P(c|\boldsymbol{x}_t, \hat{\lambda})\boldsymbol{x}_t \tag{2.14}$$

$$\boldsymbol{S}_{XX;c} = \sum_{t=1}^{T} P(c|\boldsymbol{x}_t, \hat{\lambda})\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}\,. \tag{2.15}$$

At the end of the M-step, a predetermined threshold is used to check for convergence. If convergence is not achieved, the algorithm returns to the E-step and performs another iteration, using the refined model as the initial model in the new iteration. This process is repeated until convergence is achieved.

For the diagonal covariance case, only the elements on the diagonal of $\boldsymbol{\Sigma}_c$ are retained. All off-diagonal elements are set to 0.

Although the E-M algorithm is not guaranteed to produce a globally optimal solution for the model parameters, it is guaranteed to converge to a local maximum after sufficient iterations. Since the E-M algorithm only refines the parameter estimates, an appropriate means of initialising the parameters is necessary. Reasonable initial estimates are desired as it determines which local maximum the algorithm will converge to [47], as well as affect how rapidly convergence will be achieved. A commonly used method of determining the initial estimates is the k-means algorithm [57].

2.4.2 Maximum *A Posteriori* Estimation

A more recent and commonly used approach for training GMMs is through maximum *a posteriori* (MAP) adaptation [29]. Based on Bayesian estimation theory, this approach is widely reported in speaker verification literature, such as [63]. More recently, MAP estimation of GMMs have also proven useful in training speaker cluster models for the clustering task in speaker diarization systems [5, 6, 49, 71].

Compared to the ML approach described in Section 2.4.1, the MAP approach is able to produce more robust models given limited training data, by incorporating prior knowledge about the speaker model parameters into the training procedure [78]. This is achieved through the use of a universal background model (UBM). The UBM is a GMM trained on a large selection of representative speech, often using the ML approach. Due to the large volume of training data across a large number of speakers, the UBM contains general information regarding the characteristics of speech and what a generic speaker model should look like.

For MAP estimation of a given speaker model, the prior information contained in the UBM is first acquired by initializing each model parameter according to the UBM parameters. The speaker model is then trained, or *adapted*, using utterances produced by that speaker in order to capture the speaker-specific information. Given limited speaker-specific training data, this approach enables higher order GMMs to be constructed compared to the ML approach, allowing more detailed modelling of speaker characteristics.

The ML approach described in Section 2.4.1 aims to find a set of model parameters λ^{ML} such that the likelihood of the training data is maximised, according to the criterion

$$\lambda^{ML} = \arg\max_{\lambda} p(\boldsymbol{X}|\lambda)\,. \tag{2.16}$$

From (2.16), it can be seen that the model parameter estimates depend only on the training data $\boldsymbol{X}$. On the other hand, the MAP approach constrains the model

parameters to satisfy the prior by training using the criterion

$$\lambda^{MAP} = \arg\max_{\lambda} p(\lambda|\boldsymbol{X})\,. \tag{2.17}$$

In this case, the quantity to be maximised, $p(\lambda|\boldsymbol{X})$, is the posterior probability of the model parameters after observing the training data $\boldsymbol{X}$. Applying Bayes theorem, this is equivalent to optimising

$$\lambda^{MAP} = \arg\max_{\lambda} p(\lambda)p(\boldsymbol{X}|\lambda)\,, \tag{2.18}$$

which is the likelihood of the training data multiplied by the prior distribution $p(\lambda)$. The MAP solution is therefore given by maximising $p(\lambda)p(\boldsymbol{X}|\lambda)$.

Once again, the E-M algorithm described in Section 2.4.1 can be used to solve this estimation. Following from common practice in speaker recognition and for simplicity, only the mixture component means will be adapted using prior information, while the mixture weights and covariances of the UBM are maintained. This notion is supported by the experiments presented in [63], as well as the NIST Speaker Recognition Evaluations [17].

Assuming a Gaussian prior density with component mean vectors $\boldsymbol{m}_c$, it can be shown [29] that the E-M result for the mean adaptation process is given by

$$\boldsymbol{\mu}_c = \frac{\tau_c \boldsymbol{m}_c + \boldsymbol{S}_{X;c}}{\tau_c + n_c}\,, \tag{2.19}$$

where n_c and $\boldsymbol{S}_{X;c}$ are the same statistics as used for ML estimation, and are defined in (2.13) and (2.14) respectively. It can be seen that the MAP estimation of the mean is a blend between the prior distribution mean $\boldsymbol{m}_c$ and the ML estimate based on the observed training data. The MAP solution given in (2.19) is equivalent to determining the ML solution assuming an additional τ_c samples at the prior mean $\boldsymbol{m}_c$. In the special case where $\tau_c = 0$, (2.19) reverts back to the ML solution given in (2.11). This configuration is known as a *non-informative* prior.

2.5 Eigenvoice Speaker Modelling

A more recent approach to speaker modelling, based on joint factor analysis (JFA) techniques [40], is the use of the eigenvoice model to represent the speakers [41, 44, 62]. As in traditional approaches, eigenvoice modelling techniques are based around the use of GMMs to model a speaker. Once again, only the component mean vectors will be adapted during training. The eigenvoice representation is based on the use of speaker supervectors, obtained by the concatenation of the mean vectors of each mixture component. The assumption in eigenvoice modelling is that the speaker supervectors have a Gaussian distribution of the form

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}\,, \tag{2.20}$$

where $\boldsymbol{s}$ represents a given speaker model, and $\boldsymbol{m}$ is a UBM mean supervector obtained by the concatenation of the UBM component mean vectors. $\boldsymbol{V}$ is a matrix of eigenvoices and the entries of $\boldsymbol{y}$ are the speaker factors. This is a highly informative prior distribution from a Bayesian viewpoint; as it imposes severe constraints on the speaker supervectors. Although speaker supervectors typically have a very large number of dimensions, as determined by the number of mixture components multiplied by the dimensionality of the feature vector, this representation constrains all supervectors to lie in an affine subspace of the supervector space of much lower dimension [62]. This results in a substantial reduction in the number of parameters that need to be estimated. Compared to classical MAP adaptation described in Section 2.4.2, this approach therefore requires significantly less training data for the adaptation process, which tends to saturate more quickly. However, a very large number of training speakers is required for the estimation of $\boldsymbol{V}$ [40]. More details on the eigenvoice approach to speaker modelling can be found in Chapter 6.

2.6 Speech Activity Detection

Speech activity detection, also known as speech end-point detection, is the process of classifying the audio into speech and non-speech regions. SAD is commonly used as a preprocessing step in many speech processing applications, including speaker diarization, speaker verification and speech recognition.

Traditionally, SAD has been performed by using the absolute energy as the main detection stimulus, such as in [81]. Energy based detectors are based on the assumption that for speech to exist, the time-domain signal is likely to have a burst of energy compared to non-speech regions. The raw energy values are therefore thresholded and compared against the incoming signal for speech and non-speech decisions. As a result, simple energy contour based detectors generally lack robustness against noisy environments.

In recent years, GMM based techniques [5, 35, 55, 66, 83] have received increasing attention and has become the dominant approach to the SAD task. In this approach, GMMs for different audio classes are first trained using appropriate labelled training data. In simpler SAD systems, models representing speech and non-speech classes are used [83]. More complex systems may include speech models for different gender and bandwidth conditions [55], or define specific classes such as speech over music or speech over noise, to avoid misclassifications arising due to these scenarios [66]. SAD is then performed via maximum likelihood classification using these models, commonly in conjunction with median filtering of frame scores and heuristic rules [66] to avoid rapid fluctuations between speech and non-speech classifications arising as a result of noisy likelihood scores between successive frames. Alternatively, models for the different audio classes can also be decoded using the Viterbi decoding algorithm [22, 77]. In this case, the penalty for switching between different models in the Viterbi decoding can be set to minimize unnecessary switching between different classes [5].

As stated in Section 1.2, this work will not attempt to improve current SAD techniques as it is outside the scope of this dissertation. For the purpose of evaluating speaker diarization performance, a GMM based speech activity detector

is used as a part of the baseline diarization system. This GMM based detector will remain unchanged throughout all diarization systems developed in this work, to ensure fair performance comparisons. A detailed description of the speech activity detector can be found in Section 3.5.

2.7 Speaker Segmentation

Speaker segmentation approaches reported in literature can be categorized into three classes - energy based, model selection based and metric based segmentation [13, 39].

Energy Based Segmentation

Energy based segmentation algorithms rely on silence locations within the input audio stream to partition the audio, based on the assumption that speaker segment boundaries occur at these locations. Approaches for detecting the silence locations include the use of a Viterbi decoder [35, 82], and by measuring and thresholding the signal energy [39]. The segments are then generated by partitioning the audio stream at silence locations.

In general, energy based segmentation techniques tend to perform poorly compared to both model selection based and metric based approaches [39], due to the fact that there is no direct correspondence between silence locations and changes in speaker identities. Energy based techniques are unable to detect speaker changes that occur without a corresponding silence, such as interrupted speaker segments. Moreover, short periods of silence in the middle of speaker turns are often erroneously detected as speaker change points. In [35] and [82], the decoder also provides gender information to aid segmentation, and speaker change points are assumed to occur at both silence locations and changes in the speaker's gender. However, unless separated by silence, two consecutive speakers of the same gender are still impossible to distinguish.

Similar to energy based SAD algorithms, energy based segmentation ap-

proaches also tend to lack robustness against noisy environments.

Model Selection Based Segmentation

Model selection based segmentation techniques rely on the use of competing models to determine the locations of speaker change points. In this category, by far the most popular approach reported in literature is based on the pioneering work presented in [13], which employs multivariate Gaussian modelling and uses the Bayesian information criterion (BIC) for model selection. BIC-based segmentation approaches have since received increasing attention and are widely reported in speaker diarization literature, such as [15, 66, 84].

For a speech segment with a maximum of one speaker change point, determining whether a speaker boundary is appropriate at a particular location involves the comparison of two competing models for representing the segment. One model uses a single multivariate Gaussian to represent the speech segment as a whole; while the alternative model consists of two separate Gaussian distributions representing the two speech segments, obtained by dividing the initial speech segment at the location being examined. The BIC is then used as a model selection criterion to determine whether the data in the speech segment is more appropriately modelled by one single multivariate Gaussian distribution, or two separate Gaussian distributions. If a single Gaussian is favoured, no segment boundary is assumed at the location of interest. If two separate Gaussians are more appropriate, a speaker change point is hypothesized.

In order to locate the most appropriate speaker change point within the given speech segment, every frame within the segment is examined using the above approach. The frame corresponding to the most favourable location for a speaker boundary is assumed to be the speaker change point, provided the level of dissimilarity at that point exceeds a predetermined threshold.

In segmenting real broadcast news shows with multiple speaker change points, a window-growing approach is proposed [13]. The algorithm begins with a small search window and every frame within the window is examined. If no appropriate

speaker change points are found, the window is expanded and the search process is redone. This process is repeated until an appropriate boundary is located. The search is then restarted at the location of the detected boundary, with the original small search window.

Issues in applying the BIC change detector include low detectability for boundaries between short speaker turns. The lack of data samples results in poor speaker models, and consequently genuine speaker boundaries around short speaker turns are often missed in the segmentation process. Moreover, this approach requires choosing an explicit penalty term, which is responsible for penalizing the models according to their complexities, as determined by the number of parameters in the model. In general, it is difficult to choose an appropriate penalty term that provides stable and consistent performance across different news shows, as it results in a trade-off between missed boundary detections and false alarms. For this reason, methods to adapt the penalty automatically through the use of a modified BIC criterion have been proposed [54]. Alternatively, the penalty term may be avoided altogether by controlling the model complexity [1]. More details on the BIC criterion, including its mathematical derivation, can be found in Chapter 5.

The window-growing approach is also very computationally costly, with quadratic complexity. Many systems based on this approach therefore employ some form of computation reductions, including a variable window-growing scheme [75], and the use of Hotelling's T^2-Statistic to preselect possible candidate boundaries, followed by evaluating the BIC at boundary locations to verify the hypothesized speaker change points [84]. In a similar approach referred to as DISTBIC (Distance-BIC) [15], candidate boundaries are first determined using a distance metric, and the BIC is used to either validate or discard these candidate boundaries. The use of various distance metrics are investigated, including the popular generalized likelihood ratio [30, 31] and the Kullback-Leibler divergence [70]. Distance metric based segmentation will be introduced in the following section, and is the focus of the segmentation investigations in this dissertation.

Metric Based Segmentation

In metric based segmentation, a distance metric is used to determine the statistical dissimilarity between two segments in order to make decisions as to whether a segment boundary is appropriate at a given location. This is most commonly achieved via the sliding-window approach using constant sized windows, as in [5, 70, 71]. Compared to the window-growing approach, the sliding-window approach does not involve quadratic complexity and is therefore much less computationally expensive, particularly for audio containing a large number of long speaker segments. When speaker change points are far apart, computational costs associated with the window growing approach increase significantly, whereas the costs associated with the sliding-window approach is not affected due to the constant window size.

The sliding-window approach for speaker segmentation is illustrated in Figure 2.2, where the straight vertical line at the centre of the speaker segment indicates the location of the true boundary. In this approach, a pair of adjacent sliding windows of constant length is applied across the whole audio, and the dissimilarity between the windows is evaluated using a chosen distance metric. For each frame, the level of dissimilarity at that point is determined by the distance score, obtained by evaluating the chosen distance metric using feature vectors contained within the two windows either side of that particular frame. Speaker change points are then hypothesized based on the location of the local maxima of the resultant distance curve above a predetermined threshold. Adjusting the threshold results in a trade-off between the desire to have long, pure segments to aid in initializing the subsequent clustering stage, and minimizing missed change points which result in contaminated statistics in the clustering stage [74].

To model the feature vectors contained in the adjacent windows, single multivariate Gaussians are generally preferred over GMMs, due to the simplified distance calculations. The choice of the window size is a trade-off between the need to provide sufficient data to estimate reliable speaker models, and the ability to detect boundaries between short speaker segments. Typical window sizes

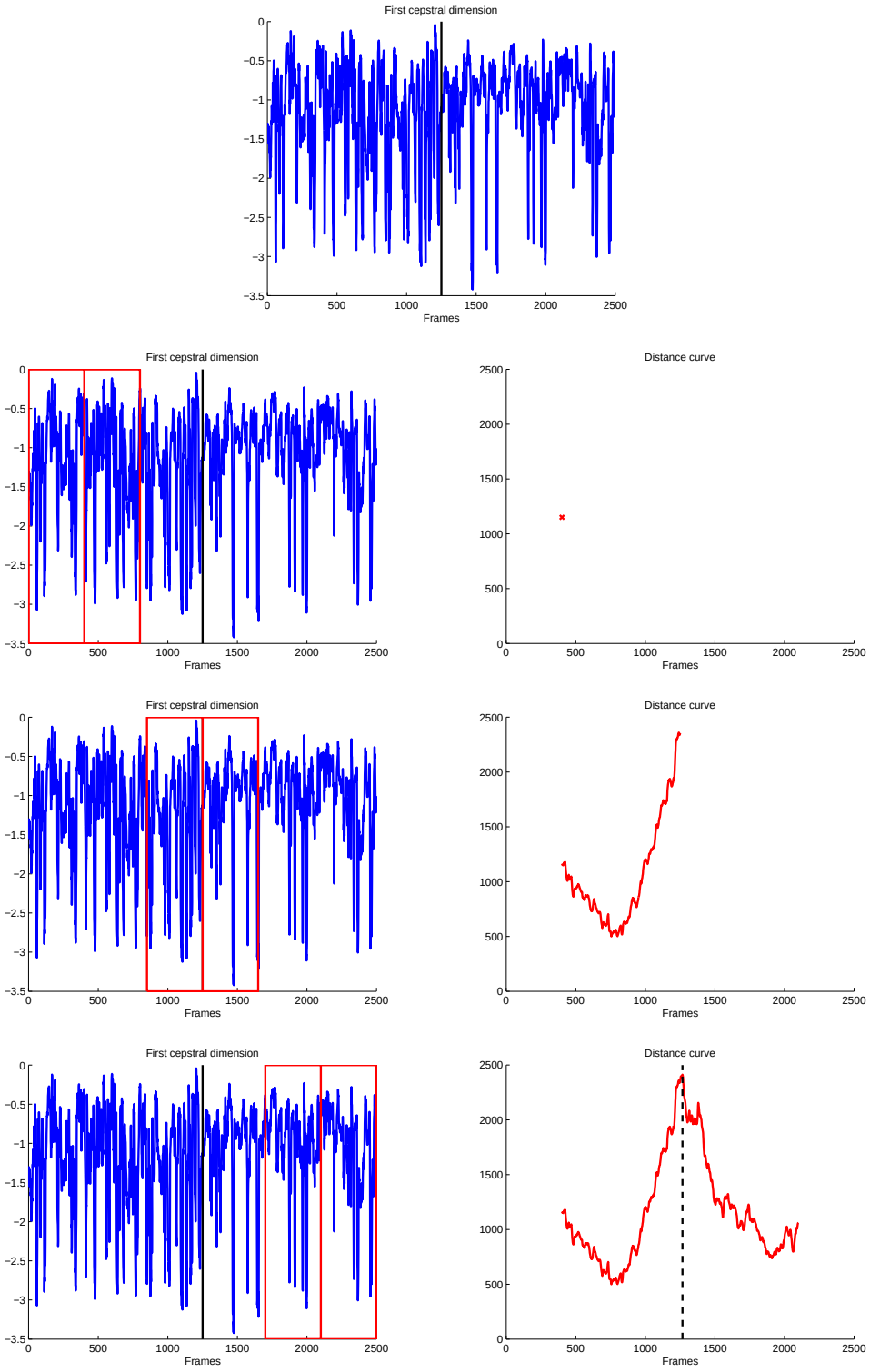

Figure 2.2: The sliding-window approach for speaker segmentation

are between 1-2 seconds when using diagonal covariance Gaussian modelling, and between 2-5 seconds when using full covariance Gaussians [74]. As with the BIC approach, the boundaries between short speaker segments are more difficult to detect, as the window length constrains the detection of short turns.

In this approach, heuristic rules are often required to govern the minimum duration constraints, smoothing of the distance curve and the elimination of insignificant local maxima points that do not correspond to true boundaries, in order to prevent the system from overgenerating change points as speaker boundaries. Heuristic rules are investigated and discussed in further detail in Chapter 4.

Various distance metrics have been reported in literature, and the most popular to date is the Kullback-Leibler (KL) divergence [70] and the generalized likelihood ratio (GLR) [30, 31]. Also referred to as the the relative cross entropy, the KL divergence is an information theoretic measure which estimates the distance between two random distributions. Since the KL divergence is not symmetric, it is not strictly a distance measure. A symmetric alternative, commonly referred to as the KL2 metric, has proved to be more popular for characterizing the similarity between two audio segments [4]. The GLR is a popular likelihood-based distance metric which has been widely reported in speaker diarization literature, such as [15, 24, 38]. For determining whether the data in the adjacent sliding windows belong to the same speaker, the GLR is given by the ratio between the likelihood of the separate models to the likelihood of the combined model. Provided that the complete statistics are available, the GLR is known as the most powerful likelihood ratio test. However, the GLR metric becomes less ideal when the data is limited, since it does not take into account the uncertainty associated with the model parameter estimates in the incomplete data case. This limitation provides a motivation towards a part of the research presented in Chapter 5. A more detailed discussion of the GLR metric can be found in Section 5.2.1.

2.8 Speaker Clustering

The speaker segmentation stage ideally produces pure, homogeneous speaker segments containing one speaker each. The clustering stage is then responsible for associating the segments belonging to the same speaker together. Ideally, at the end of the clustering stage, one cluster containing all segments from a given speaker will be produced for each speaker in the audio. This general approach to the speaker diarization problem, namely speaker segmentation followed by clustering, is referred to as the *bottom-up* approach and is the most popular approach reported in literature. An alternative approach is known as the *top-down* approach, where the entire audio stream is first modelled with a single speaker model, and new models are successively added until the full number of speakers are deemed to be accounted for. As this approach is far less popular than its bottom-up counterpart and is generally outperformed by the best bottom-up approaches [4], it will not be covered in this work. Examples of top-down approaches to speaker diarization can be found in [9, 48].

For bottom-up approaches, the speaker clustering stage is performed via agglomerative hierarchical clustering. In this approach, speaker segments produced by the segmentation stage are used to seed one initial cluster each. Clustering is then performed by modelling each cluster and evaluating the pairwise similarity between all pairs of clusters, and iteratively merging the closest pair of clusters until no more merge candidates can be found.

As with segmentation techniques, the similarity between clusters are determined by evaluating a chosen distance measure, and the choice of an appropriate distance measure is essential to the success of the clustering system. In general, distance metrics used for segmentation can also be used for clustering, although the BIC is the predominant approach for single, full covariance Gaussian modelling [13, 51, 66] and the cross likelihood ratio (CLR) is often adopted for GMM based modelling [5, 71].

The use of the BIC for agglomerative clustering is conceptually similar to segmentation. Recall from Section 2.7, the BIC can be used as a model selec-

tion criterion to determine whether the data contained in the two clusters of interest are more appropriately represented by one combined model, or two separate models. For the clustering application, if the combined model is favoured, this indicates that the two clusters should be merged. If two separate models are deemed more appropriate, the clusters should be kept separate. Similar to its segmentation counterpart, variations of BIC based techniques have also been proposed in literature for the clustering task. In [2] and [83], a modified BIC criterion is employed, which eliminates the need for tuning the BIC penalty weight by ensuring that the number of parameters in the combined model is equal to the sum of the number of parameters in the separate models. Alternatives to the penalty term, such as using a constant [73], or the weighted sum of the number of clusters and number of segments [28], have also achieved reasonable success, however the BIC method has generally outperformed those approaches [74]. The addition of a Viterbi resegmentation stage, which aims to improve diarization performance by refining the speaker boundary locations, have also been reported in literature. Viterbi resegmentation can either be performed between multiple iterations of clustering [5] or within a single iteration [83].

Despite its popularity, the derivation of the BIC involves some approximations which result in the prior term being omitted [80]. The BIC is therefore not strictly a *Bayesian* criterion, in the sense that it does not require or take into account any prior information regarding the speaker segments. Chapter 5 provides a detailed discussion of this issue, including a mathematical derivation of the BIC criterion.

2.9 Speaker Diarization Using Eigenvoice Modelling Techniques

The difficulty with estimating reliable models for short speaker segments, due to the lack of data, is a well-known limitation of GMM or single Gaussian based modelling approaches to speaker diarization [13, 15, 84]. The modelling of speaker segments using the eigenvoice framework, introduced in Section 2.5, is able to

exploit the highly informative prior knowledge about the speaker space to find a low dimensional vector of speaker factors that summarize the salient speaker characteristics, by constraining the speaker models to a previously estimated linear subspace. These constraints allow a reliable speaker model to be estimated using limited data.

The factor analysis approach to speaker modelling have become increasingly popular in recent speaker recognition literature. While earlier work in this area have primarily focused on the speaker verification task, such as in [43] and [79], increasing research efforts are being placed on the application of eigenvoice modelling techniques in the speaker diarization task in recent years. Speaker diarization using eigenvoice modelling was first introduced in [12]. This system performs an online diarization where the incoming audio stream is treated as a stream of fixed-duration time slices. Segmentation and clustering is then performed in a causal manner, that is, an incoming audio slice is processed on the fly without requiring the following slices. For this approach, the detection of the speaker turns and the estimation of the speaker models therefore require low complexity in order to cope with audio streaming. Given an audio slice, a stream of speaker factors is computed using a small, symmetric sliding window of about 1 second in length over each frame of the audio slice. Segmentation and clustering is then performed using these speaker factors.

A non-causal diarization system using eigenvoice modelling is the Variational Bayes system, as reported in [44, 62]. Inspired by the pioneering work by Valente [76], which uses probabilistic methods for speaker clustering and invokes Variational Bayesian techniques as an approximate inference method for approximating intractable integrals and posteriors, this system replaces the prior distribution on GMMs used by Valente in constructing the hierarchical generative model, with the stronger eigenvoice and eigenchannel priors used in factor analysis based speaker recognition. This led to superior results in the diarization of telephone conversations compared to the stream-based approach [44]. The complete mathematical derivations behind this approach can be found in [41].

2.10 Summary

This chapter presented a review of recent approaches to the three main stages within a speaker diarization system, namely speech activity detection, speaker segmentation and speaker clustering. Much of this review was devoted to techniques based on traditional statistical pattern recognition approaches to speaker modelling, using single multivariate Gaussian or GMM modelling of features extracted from short-time analysis of the spectral content of speech signals.

Recent developments in speaker diarization techniques using eigenvoice modelling of speaker segments were also described. The advantages of eigenvoice modelling over traditional GMM based approaches were discussed, with emphasis placed on its ability to exploit the highly informative prior information about what speaker models should look like, thus allowing the estimation of reliable speaker models using limited training data.

Various approaches to speaker segmentation strategies were outlined, using traditional approaches to speaker modelling. Popular speaker segmentation strategies were detailed, and the sliding-window approach was presented as an effective method of locating speaker change points without the computational demands associated with the window-growing approach. Popular distance metrics reported in literature were also discussed. Of particular interest is the GLR, which is known as the most powerful likelihood ratio test, provided that the complete statistics are available. The difficulties with estimating reliable models for short speaker segments using traditional modelling approaches were also highlighted. This limitation results in a higher number of missed detections for boundaries around short speaker segments.

The speaker clustering task is conceptually similar to its segmentation counterpart, in terms of evaluating the similarity between two given speaker segments. In general, all distance metrics used for segmentation can also be employed for clustering, although the application of the BIC model selection criterion in an agglomerative hierarchical clustering framework using single multivariate Gaussian modelling was presented as the predominant approach for this task.

Finally, recent approaches to speaker diarization using eigenvoice modelling techniques were outlined. The ability to estimate reliable speaker models based on limited training data allows improved modelling of short speaker segments compared to traditional modelling techniques. Eigenvoice modelling for speaker diarization forms a significant part of the research presented in this dissertation, and is detailed in Chapter 6.

Chapter 3

Speaker Diarization Evaluation

3.1 Introduction

In order to evaluate and compare the performance of different techniques used in speaker diarization systems, suitable performance evaluation metrics and testing protocols are required. Appropriate databases containing labelled speech audio are also essential for the training of relevant speaker models and to provide testing data for the evaluation of diarization performance.

This chapter outlines the standardized performance evaluation metrics and testing protocols as used in the speaker diarization task of the U.S. National Institute of Standards and Technology (NIST) Rich Transcription (RT) Evaluations [19]. As stated in Section 1.2, the scope of this research programme is restricted to speaker segmentation and clustering on the broadcast news domain. Databases used throughout this dissertation are therefore also restricted to this genre.

Following the description of performance evaluation metrics, testing protocols and databases, Section 3.5 describes in full the baseline system used throughout this dissertation for performance comparisons. The baseline system is based on the *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur* (LIMSI) broadcast news diarization system [5], which was the top participant in the speaker diarization task of the NIST Fall 2004 Rich Transcription (RT-04F)

Evaluation [19].

3.2 Performance Evaluation Metrics

As the research presented in this dissertation focuses on improving speaker diarization performance, all systems developed throughout this dissertation will be evaluated using the diarization error rate (DER) measure, as defined in [19]. The DER is the primary performance evaluation metric used in the NIST RT diarization tasks and, likewise, will also be used as the primary evaluation metric throughout this dissertation. Other performance metrics described in this section are used as an indication of the performance of the relevant components within the speaker diarization system, namely the speaker segmentation and clustering stages.

Diarization Error Rate

The DER is a time-based measure which can be interpreted as the proportion of the total amount of scorable time that is not attributed to the correct speaker, taking into account speech detection errors. It is calculated via an optimal one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs so as to maximize the total overlap between the reference and mapped hypothesis speakers. The total speaker diarization error time is given by the sum of the miss (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference) and speaker error (mapped reference speaker is not the same as the hypothesized speaker) times.

As stated in [19], the DER can be formally expressed as

$$Error_{SpkrSeg} = \frac{\sum_{\text{all segs}} \left\{ \text{dur}(seg) \cdot \left(\max\left(N_{Ref}(seg), N_{Sys}(seg)\right) - N_{Correct}(seg) \right) \right\}}{\sum_{\text{all segs}} \left\{ \text{dur}(seg) \cdot N_{Ref}(seg) \right\}}, \tag{3.1}$$

where the speech data file is divided into contiguous segments at all speaker

change points and where, for each segment, seg,

$$\begin{aligned}
\text{dur}(seg) &= \text{the duration of } seg, \\
N_{Ref}(seg) &= \text{the number of reference speakers speaking in } seg, \\
N_{Sys}(seg) &= \text{the number of system speakers speaking in } seg, \\
N_{Correct}(seg) &= \text{the number of reference speakers speaking in } seg \text{ for whom} \\
&\quad \text{their matching (mapped) system speakers are also speaking} \\
&\quad \text{in } seg.
\end{aligned}$$

Speaker Segmentation Evaluation Metrics

The performance of the speaker segmentation stage can be evaluated by the miss and false alarm rates, as in [13]. A missed detection is defined as a segment boundary in the reference ground-truth which the system failed to detect. A false alarm is defined as an erroneously detected boundary which fails to match a corresponding boundary in the ground-truth. The miss and false alarm rates are given by

$$\text{miss rate} = \frac{\text{number of missed detections}}{\text{total number of true boundaries}} \tag{3.2}$$

$$\text{false alarm rate} = \frac{\text{number of false alarms}}{\text{total number of detected boundaries}} . \tag{3.3}$$

Speaker Clustering Evaluation Metrics

The performance of the speaker clustering stage can be evaluated by the time-based cluster purity and cluster coverage measures, as defined in [28]. For a given cluster, cluster purity is defined as the ratio of the number of frames belonging to the most dominant speaker in the cluster, to the total number of frames in the cluster. Cluster coverage accounts for the dispersion of a given speaker's data across clusters. For a given speaker, cluster coverage is defined as the ratio of the number of frames belonging to the given speaker within the cluster which has most of the speaker's data, to the total number of frames belonging to that speaker across all clusters.

In order to calculate the average cluster purity for a given audio, the cluster purities of all clusters are first calculated. The average cluster purity is then given

by the average purity across all clusters, time-weighted according to the number of frames in each cluster. Similarly, the average cluster coverage for a given audio is given by the average coverage across all speakers, time-weighted according to the number of frames belonging to each speaker.

3.3 Databases

This section outlines the databases used throughout this research programme. The NIST 2002 Rich Transcription Evaluation Data is used to evaluate the performance of diarization systems developed throughout this dissertation. All other databases described in this section are used for training the relevant speaker models used in various stages within the developed systems. All recordings in all databases are in English.

NIST 2002 Rich Transcription Evaluation Data

The NIST RT-02 Evaluation Data [26, 27] was collected by the Linguistic Data Consortium (LDC) as part of the Effective, Affordable, Reusable Speech-to-text (EARS) project. This dataset contains audio in the Broadcast News and Conversational Telephone Speech domains, however only Broadcast News audio will be used throughout this work. The Broadcast News data consists of six news shows from six different television and radio broadcasts, namely MNB, PRI, NBC, CNN, VOA and ABC, all collected in 1998. All Broadcast News shows contained in this dataset are previously unreleased by the LDC. From each of the six shows, ranging from 30 to 60 minutes each, a ten minute excerpt is chosen as the scorable region, where speaker segments are labelled with the speaker identities and their corresponding start and end times.

Hub-4 English Broadcast News Corpora

The LDC began its first Broadcast News speech collection in 1996. The primary motivation for the collection is to provide training data for the U.S. Defense Ad-

vanced Research Projects Agency (DARPA) 'Hub-4' Project, which focuses on continuous speech recognition in the Broadcast News domain. The 1996 Hub-4 Broadcast News corpus [34] contains a total of 104 hours of broadcasts from ABC, CNN and CSPAN television networks and NPR and PRI radio networks. All 174 recordings were collected in 1996, and range from 30 to 120 minutes. Transcripts are provided for all recordings, manually time aligned to the phrasal level, annotated to identify boundaries between news stories, speaker turn boundaries, and gender information about the speakers. A number of conditions, including dialect (native or nonnative), mode of speech (planned or spontaneous), fidelity (high, medium or low) and background (clean, music, speech or other noise) are also identified and timesegmented [25].

The 1997 Hub-4 Broadcast News corpus [21], collected between 1997 and 1998, was designed to serve as a supplement to the 1996 collection, providing an additional 97 hours of training data for the DARPA Hub-4 Project. The 1997 collection, containing 114 recordings, is sourced from ABC, CNN, CSPAN television networks and the PRI radio network. Transcripts for all recordings are provided in a comparable format to those used in the 1996 collection.

More details of the Hub-4 Broadcast News corpora, including the breakdown of each collection into its specific sources, can be found in [33].

USC Marketplace Broadcast News Corpus

The USC Marketplace Broadcast News Corpus [11] contains approximately 40 hours of radio broadcast news data, recorded in 1996. The audio consists of 84 recordings, none of which are already included in the Hub-4 collections. Transcripts were created using Hub-4 style conventions, and contain information including story boundaries, disfluency markers, and speaker and gender identification. All speaker segments are labelled with their corresponding start and end times. Commercial and music segments, while a part of the audio publication, are excluded from the transcripts.

3.4 Testing Protocols

This section outlines the testing protocols used throughout this dissertation for evaluating the performance of the developed diarization systems, in accordance with the strict testing protocols used in the NIST RT evaluations [19].

All diarization systems developed throughout this work hypothesize a set of relative speaker segments, each consisting of a speaker-ID label and the corresponding start and end times. In order to measure the performance of each system, the speaker labels are scored against a reference ground-truth via an optimum one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs so as to maximize the total overlap of the reference and mapped hypothesis speakers. Performance of the system is then evaluated in terms of the miss, false alarm, and speaker-error rates that result from the mapping, using the DER metric as defined in Section 3.2. Should overlapping regions of speech exist, the speaker mapping will be computed over all speech including regions of overlap, however the DER will be computed over only the non-overlapping speech. A time collar of 0.25 seconds each side of any reference speaker boundary will not be scored, which allows for timing errors in the reference.

For each speech data file, mapping of speakers is computed independently, and the DER is also evaluated separately using the relevant ground-truth and optimum mapping of speakers. The average DER is then calculated based on a time weighted average according to the amount of scorable time in each show. All speaker labels produced by the diarization system outside the scorable region are ignored in the scoring process.

3.5 Baseline Speaker Diarization System

This section describes the baseline system used throughout this dissertation for performance comparisons. The baseline system is based on the *c-sid* configuration of the LIMSI Broadcast News diarization system [5], which was the top participant in the most recent NIST Rich Transcription Broadcast News evalu-

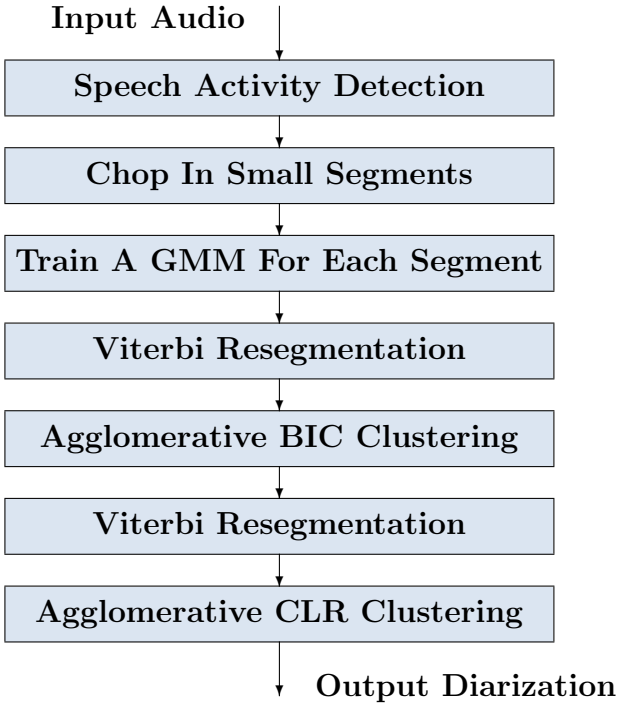

Figure 3.1: Baseline speaker diarization system

ation, the RT-04F [19, 20]. The baseline speaker diarization procedure is shown in Figure 3.1. The audio features extracted for each stage of the system is based on the features used in [5]. Each stage of the baseline system is as follows:

Speech Activity Detection: The speech activity detection (SAD) stage is responsible for classifying the audio into speech and non-speech regions. Speech is extracted from the signal via Viterbi decoding using 64-mixture Gaussian Mixture Models (GMM) with diagonal covariance. As opposed to using a simple two-class speech/non-speech detector, such as in [83], five class models are used in this stage. The five classes in the classifier include three speech classes (speech, speech+music, and speech+noise) and two non-speech classes (music and noise), as in [66]. GMMs representing each of the five classes are trained using approximately 1 hour of the specific type of data, selected from the 1996 and 1997 Hub-4 Broadcast News corpora. The 'speech+music' and 'speech+noise' classes are used to help minimize the false rejection of speech in the presence of music or noise, and frames that are identified as belonging to these classes are subsequently

reclassified as speech. The aim of this stage is to remove only 'significant' non-speech regions. The penalty of switching between models in the Viterbi decoding is therefore set to minimize unnecessary switching between different classes. This prevents short pauses in the middle of speaker turns to be classified as non-speech, thus breaking up homogeneous speaker segments. All audio regions classified as non-speech are excluded from further processing.

Chop In Small Segments: This is the speaker segmentation stage, responsible for partitioning the speech regions into homogeneous speaker segments. This is achieved through the sliding-window approach [70], where a pair of adjacent sliding windows, each containing 500 frames representing 5 seconds of speech, is applied across the whole audio. For each frame throughout the audio, the generalized likelihood ratio (GLR) is calculated to evaluate the level of dissimilarity between the feature vectors contained within the two windows either side of that particular frame. Local maxima of the resultant distance curve above a predetermined threshold are then assumed to be speaker change points. A minimum duration constraint of 2.5 seconds is placed on all speech segments. Should two local maxima of the distance curve above the threshold exist within a 2.5 second proximity, the segment boundary corresponding to the local maximum point with a lower GLR score is rejected. The segmentation threshold is chosen to minimize missed boundary detections at the expense of higher false alarms, since false change points can be removed later during the clustering process.

Train A GMM For Each Segment: For each initial segment, a 4-mixture GMM with diagonal covariance matrix is trained using data contained in the segment. The GMMs trained on each initial segment are then passed into the Viterbi Resegmentation stage.

Viterbi Resegmentation: The Viterbi Resegmentation stage aims to refine the segment boundary locations by performing a forced alignment using GMMs trained from each segment. Each refined segment is then used to seed one initial cluster.

Agglomerative BIC Clustering: This is the first of two speaker clustering

stages. In this stage, the Bayesian Information Criterion (BIC) is used as a decision criterion for agglomerative clustering, as in [13]. As speaker segments are still relatively short in this initial clustering stage, a single multivariate Gaussian distribution with full covariance matrix is used to model the data, as opposed to a GMM. This initial clustering stage merges only the closest speaker segments and is terminated early, resulting in a set of underclustered nodes. The performance of this clustering stage is crucial to the success of the overall diarization system, since correct clustering decisions made in this stage will generate pure, homogeneous clusters with sufficient data to be represented by more complex models in the subsequent clustering stage. On the other hand, incorrect clustering decisions will result in impure clusters with contaminated statistics, deteriorating subsequent clustering performance. At the end of the initial clustering stage, a GMM is trained for each cluster and the segment boundaries are refined once more via Viterbi resegmentation.

Agglomerative CLR Clustering: This is the second clustering stage, which completes the clustering process using the cross likelihood ratio (CLR) criterion, as in [65]. In this clustering stage, the initial clusters have considerably more data than the individual speaker segments passed into the first clustering stage. GMMs are therefore used to model the more complex distributions of data in each speaker cluster. A show background model, represented by a 128-mixture GMM, is first trained using all speech segments from the whole show. Models for each individual speaker cluster are then obtained via Maximum *a posteriori* (MAP) adaptation of the GMM means from the show background model, using data from the relevant cluster of interest. The CLR between each pair of clusters is then calculated and agglomerative clustering is performed. After each merge, a new GMM is trained for the merged cluster via MAP adaptation from the show background model, using data from both merge candidates. The closest pair of clusters are merged iteratively until no more suitable merge candidates can be found. The second clustering stage produces the final diarization output, consisting of a relative, show-internal set of speaker labels and their corresponding

start and end times.

3.6 Summary

This chapter outlined the Broadcast News speech corpora used to provide training data for the development of diarization systems throughout this dissertation, as well as evaluating the performance of the developed systems. Standardized performance evaluation metrics and testing protocols used throughout this work, in accordance with the NIST RT Evaluations, are also reviewed.

Finally, a description of the baseline Broadcast News diarization system was presented, detailing each stage of the diarization procedure. This system is used throughout Chapters 4, 5 and 6 for performance comparisons.

Chapter 4

Speaker Segmentation Heuristics

4.1 Introduction

Speaker segmentation, the process of partitioning the audio data into homogeneous segments according to speaker identities, is often performed as one of the first stages within a speaker diarization system [74]. The segmentation stage is responsible for determining all boundary locations within a given audio that correspond to true speaker change points, providing clean, uncontaminated data for subsequent speaker clustering. Speaker segmentation is therefore commonly regarded as the most crucial step in the first stages of a speaker diarization system.

Recall from Section 2.7, one of the most popular speaker segmentation strategies to date uses a sliding-window approach, where a pair of adjacent sliding windows of constant length is applied across the whole audio, and the dissimilarity between the windows is evaluated using a chosen distance metric. Local maxima of the resultant distance curve with distance scores exceeding a predetermined threshold are then used to locate the speaker boundaries. In this approach, determining the threshold results in a trade-off between missed boundaries and false alarms. A large threshold generally corresponds to a high miss rate and a low false alarm rate, and vice versa. In speaker diarization applications, it is advantageous to minimize missed boundaries at the expense of high false alarms at this stage; as long as the speaker segments are of sufficient duration in order to

characterize the voice of the speakers [5]. Erroneously detected boundaries can be rectified in the subsequent speaker clustering process. On the other hand, missed boundaries are difficult to rectify, thus resulting in speech from multiple speakers being assigned to the same speaker segment. This contamination of statistics in turn deteriorates subsequent clustering performance.

This chapter aims to outline some initial investigations into strategies of improving speaker segmentation performance, with the emphasis of minimizing missed boundary detections. Due to the nature of speech signals, the resultant distance curve is intrinsically noisy, and not all local maxima of the distance curve above a given threshold correspond to true speaker change points. The following sections describe how heuristic rules can be applied to determine which local maxima correspond to true segment boundaries, as well as demonstrating the advantages of the heuristic approach over a single threshold in minimizing missed boundary detections. The generalized likelihood ratio (GLR) distance metric is used throughout this chapter, as in the baseline system described in Section 3.5.

This chapter is structured as follows: Section 4.2 illustrates the noisy nature of the GLR distance curve, which motivates the development of the heuristic segmentation algorithm proposed in this work. Section 4.3 presents the proposed algorithm and details the techniques behind each step of the development of the heuristic process. Section 4.4 outlines the experiments performed and results obtained on the Rich Transcription 2002 (RT-02) Evaluation Dataset and compares the outcome of the proposed segmentation strategy to the baseline algorithm. Section 4.5 provides some discussions regarding the robustness of the heuristic technique compared to the single threshold based segmentation strategy used in the baseline system.

4.2 The Noisy Nature of Distance Curves

As with any distance metric, the GLR distance curve is intrinsically noisy. Figure 4.1 shows a typical GLR curve in the absence of any speaker segment boundaries.

An example of a typical GLR curve with speaker change points is given in Figure 4.2, where the straight vertical lines indicate the true boundary locations. From Figure 4.2, it can be observed that speaker boundaries generally occur at the location of local maxima points (referred to as 'peaks' in the remainder of this chapter) with large GLR values within the distance curve, although not all such peaks correspond to true speaker change points. The noisy nature of the GLR distance curve results in many false peaks, with both high and low GLR scores, that do not correspond to true segment boundaries. The use of a single threshold, as in the baseline system described in Section 3.5, is therefore insufficient to ensure accurate segmentation. The next section presents the heuristic rules developed in this research, aimed to identify which peaks correspond to true speaker boundaries.

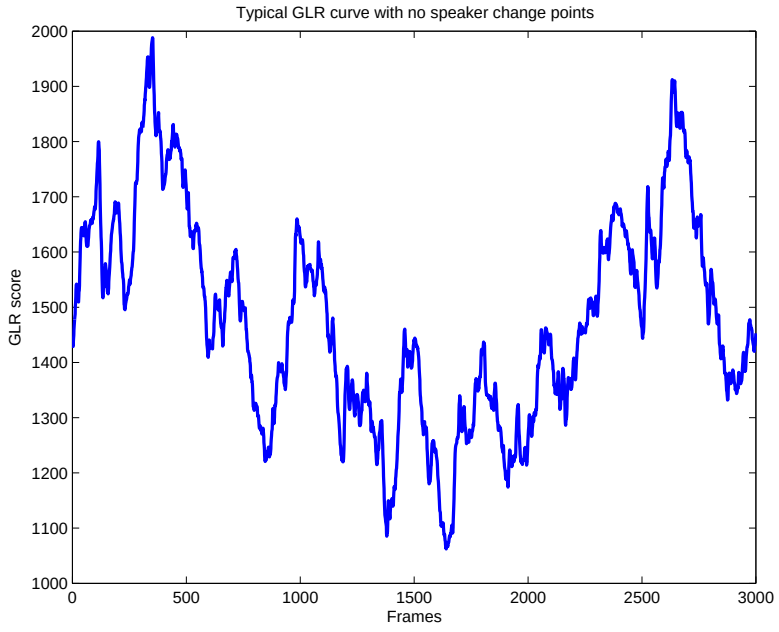

Figure 4.1: Typical GLR curve in the absence of speaker segment boundaries

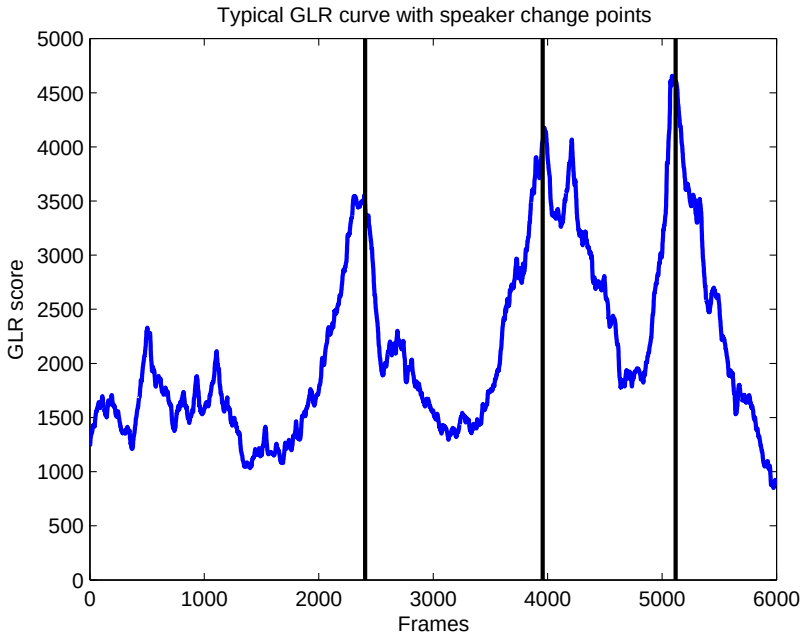

Figure 4.2: Typical GLR curve with speaker segment boundaries

4.3 Heuristic Rules for Speaker Segmentation

This section outlines the heuristic rules developed in this research aimed to identify which peaks correspond to true speaker change points, with the emphasis of minimizing missed boundary detections. The techniques used to develop the heuristic algorithm proposed in this work are detailed in Sections 4.3.1 to 4.3.3. The proposed algorithm is summarized in Section 4.3.4.

4.3.1 Determining Peak Locations

Before any GLR scores can be computed, the size of the adjacent sliding windows must be chosen. Large windows provide improved accuracy for detecting exact boundary locations between longer speaker segments, due to the abundance of available data for calculating the GLR score. On the other hand, smaller windows offer improved detectability of boundaries around short speaker segments, although they are also more prone to false alarms arising in the middle of speaker turns that do not correspond to true speaker change points. Adjusting the win-

dow size therefore results in a trade-off between the need for accurate boundary detections, and the desire to minimize missed detections at the expense of increased false alarms.

Once the appropriate window size is chosen, the GLR score between the pair of adjacent sliding windows is computed over the entire audio. Due to the fact that the length of a single speech frame is insignificant compared to the length of a typical speaker segment, not all points along the GLR distance curve need to be considered for further processing. Extracting every 10th sample from the distance curve significantly reduces the computational demand of subsequent steps, as well as achieving a smoothing effect. This smoothing effect greatly reduces the number of local false peaks consisting of only a few frames, whilst maintaining the 'significant' peaks that are essential for the accurate detection of true segment boundaries. Further smoothing is achieved by the application of a median filter over the smoothed GLR curve.

The locations of the peaks are determined by computing the first derivative of the smoothed GLR curve, and finding the locations of its zero crossings changing from positive to negative. Each peak is then examined, 'significant' peaks are assumed to be speaker change points, and all other peaks are ignored. Heuristic rules used to distinguish 'significant' peaks from false peaks are detailed in the next section.

4.3.2 Determining Significant Peaks

Although not all peaks with large GLR values correspond to true speaker change points, peaks with low GLR scores are generally false peaks. Therefore, all peaks with scores that fall below a predetermined threshold, tuned to minimize missed detections, are assumed to be false peaks and are discarded without further processing.

The fluctuations of the GLR value above the segmentation threshold that has not been smoothed out by the sampling or median filtering processes create false peaks with GLR scores above the threshold. These false peaks cannot be easily

removed via a simple comparison against a threshold, since their GLR values are comparable to that of true segment boundaries. One distinctive characteristic which can be used to identify these false peaks is the lack of depth of the local minima (referred to as 'troughs' in the remainder of this chapter) between the false peak and its neighbouring peak. An example of this is shown in Figure 4.2, where the shallow trough between the two peaks, located at approximately 4000 and 4200 frames respectively, suggest that at least one of the two peaks is a false peak. In this case, the shorter of the two peaks, located at approximately 4200 frames, is a false peak. The removal of such false peaks aims to provide improved segmentation performance by eliminating false boundaries with large GLR values.

4.3.3 Segmentation Incorporating a Smaller Window

As discussed previously, while large windows have the advantage of improved accuracy of boundary detections around longer speaker segments, smaller windows have the advantage of improved detectability of boundaries between short segments. This is illustrated in Figure 4.3, where the resultant GLR curve computed using smaller windows provides a stronger indication of where the true boundaries should be. It is therefore possible to capitalize on the advantages of different sized windows by combining the information given, in order to create a segmentation system that is accurate in locating the speaker boundaries around long segments, as well as being capable of detecting boundaries around short segments. The proposed heuristic algorithm is summarized in the next section.

4.3.4 Proposed Heuristic Algorithm

The techniques presented in Sections 4.3.1 to 4.3.3 motivate the following heuristic segmentation algorithm. The GLR score between a pair of large adjacent windows is first calculated over the entire audio. This is followed by sampling and median filtering, after which the locations of all peaks are identified using the process described in Section 4.3.1. The GLR values of all peaks are then compared

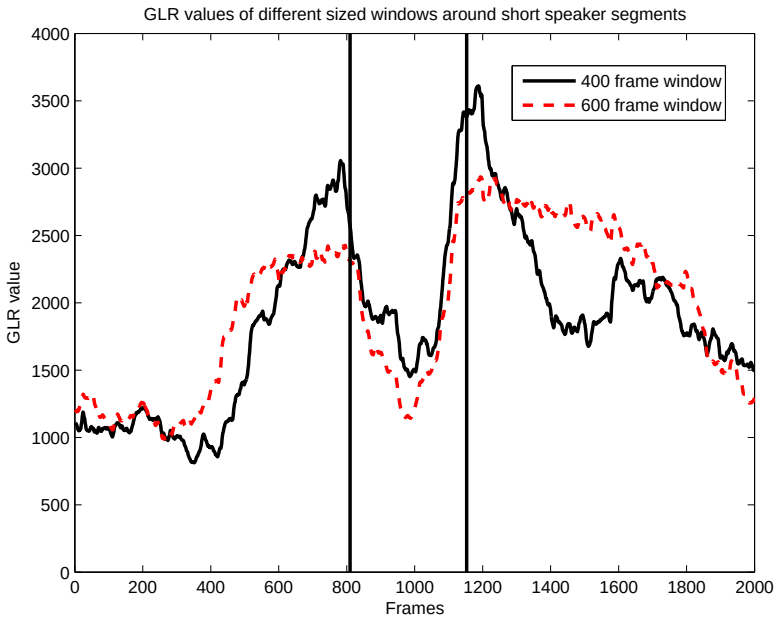

Figure 4.3: Comparison of different window sizes and its effects on boundary detection around short segments

against a predetermined segmentation threshold, and all peaks that fall below the threshold are discarded without further processing. Next, all false peaks with GLR values above the threshold are identified by analyzing the depth of the troughs shared with their neighbouring peaks. If the GLR ratio between the trough and the shorter neighbouring peak is greater than a predetermined threshold, this indicates that at least one of the two peaks either side of the trough is a false peak. The shorter peak is then assumed to be a false peak and is discarded. Note that the taller of the two peaks is not guaranteed to be a true peak; its acceptance as a true segment boundary will depend on the depth of the trough shared with its next neighbouring peak.

Since large windows are used, the set of speaker change points produced by this heuristic process is likely to provide accurate detection of boundaries around long speaker segments. This process is then repeated using a smaller window, with a different segmentation threshold, to produce a second set of speaker change points with improved boundary detection around shorter segments. The two sets of speaker change points are then merged to generate the final segmentation. All

boundaries detected using the large windows are honoured. The boundaries detected by the smaller windows are then examined in turn; if a detected boundary lies within close proximity to an existing boundary found by the larger window, the existing boundary is assumed to be more accurate and the detected boundary is therefore ignored. If no boundaries already exist within close proximity, the detected boundary is honoured and added to the final segmentation output.

The proposed heuristic algorithm detailed in this section is depicted in Figure 4.4. All parameters used in the heuristic algorithm, which are emperically determined for optimal performance, are summarized in Table 4.1.

Table 4.1: Summary of Parameters and their Emperically Determined Values

Parameter	Value
Large Window Size (frames)	600
GLR Curve Sampling Interval (frames)	10
Median Filter Length	7
Trough to Peak GLR Ratio for Determining False Peaks	>0.8
Small Window Size (frames)	400
Difference in Segmentation Threshold between Large and Small Windows	500
Proximity to Existing Boundary for Rejection of Detected Boundary (frames)	<150

4.4 Experiments

This section presents the segmentation and diarization results obtained using the parameter values shown in Table 4.1 over a range of segmentation thresholds, and compares the results to the baseline system. The baseline system used in this study is described in Section 3.5.

4.4.1 Segmentation Results

The segmentation results presented in this section is calculated using all 6 shows in the RT-02 Evaluation Dataset. For each system, the segmentation results are presented across a range of segmentation thresholds. For each threshold, the total number of missed boundary detections and false alarms across all 6 shows

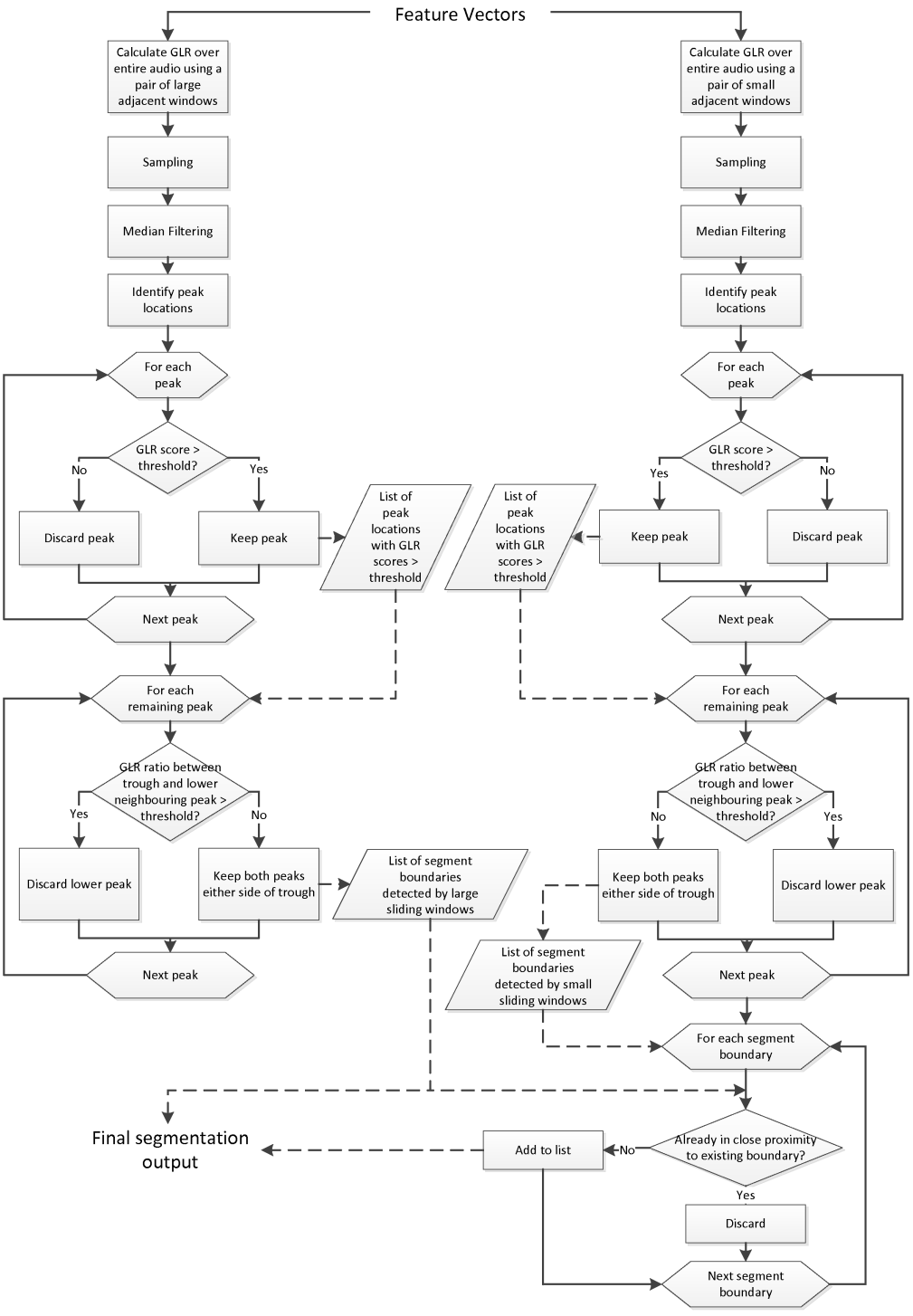

Figure 4.4: Proposed heuristic algorithm

are used to calculate the miss and false alarm rates, which are defined in Section 3.2. Further details of the RT-02 Evaluation Dataset can be found in Section 3.3.

Figure 4.5 shows a comparison of the overall 'error score' achieved by each system across a range of segmentation thresholds used by the larger window. As presented in Table 4.1, the segmentation threshold used for the smaller window is 500 less than the larger window, a value emperically determined for optimal performance. The error score is defined as the sum of the miss and false alarm rates. Of particular interest in speaker diarization applications is the low threshold region, which is associated with lower miss rates and higher false alarm rates. Improved segmentation performance is attained with the heuristic method not only in the low threshold region, but across a wide range of thresholds.

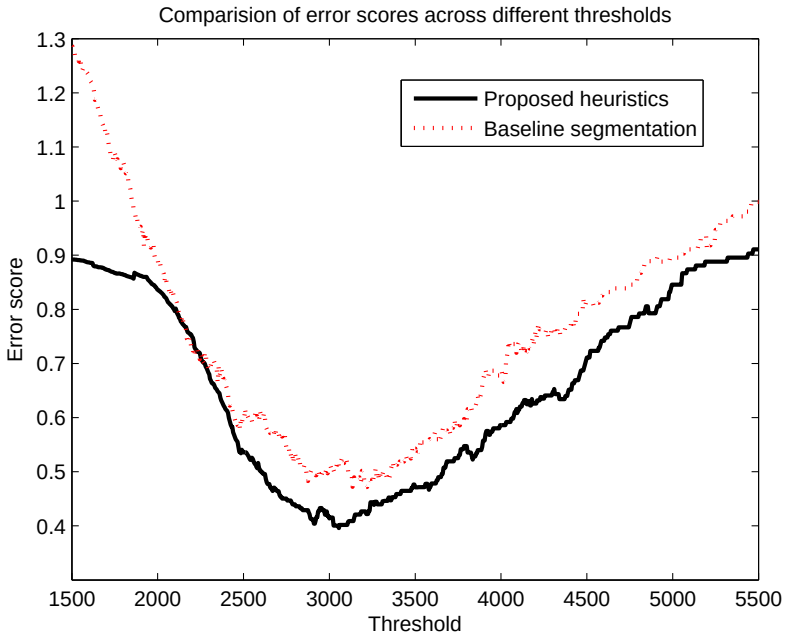

Figure 4.5: Comparison of error scores between baseline and proposed heuristic segmentation

Figures 4.6 and 4.7 show a breakdown of the error score into its respective components. It is interesting to note that most of the improvement in segmentation performance obtained by the heuristic system comes from the reduction in miss rates, while false alarm rates are more similar across all thresholds. This is

a desirable attribute in speaker diarization applications, as false alarms can be corrected in the subsequent clustering stage, while missed boundary detections are difficult to rectify.

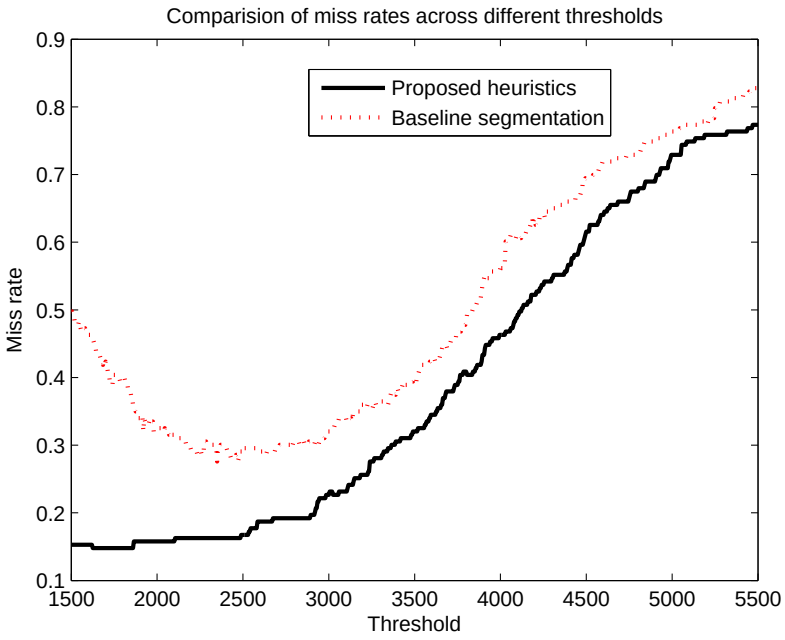

Figure 4.6: Comparison of miss rates between baseline and proposed heuristic segmentation

4.4.2 Diarization Results

The overall diarization performance of the two systems, evaluated using the diarization error rate (DER) measure defined in Section 3.2, is shown in Table 4.2. The average DER reported is time weighted according to the length of the scorable region within each individual show in the RT-02 Evaluation Dataset. The proposed heuristic segmentation system is simply the baseline diarization system with the segmentation stage replaced by the proposed heuristic algorithm. These results show the effect of improved segmentation on overall diarization performance. The average DER shown are obtained by using the optimal local stopping threshold for each show in the final clustering stage. As evident from Table 4.2, an overall relative improvement in DER of 4.7 % is achieved when using heuristic

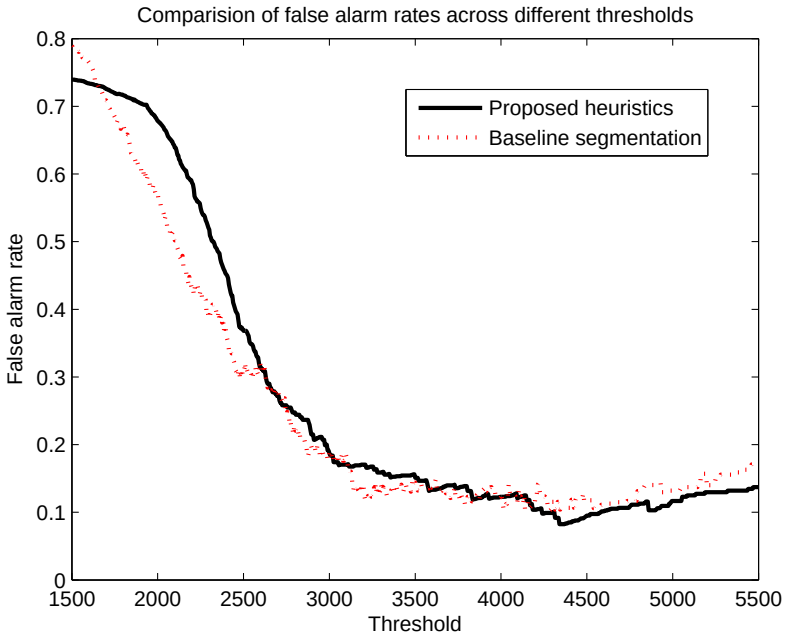

Figure 4.7: Comparison of false alarm rates between baseline and proposed heuristic segmentation

segmentation.

Table 4.2: Comparison of overall diarization error rates (%)

System	Average DER
Baseline	10.6
Heuristic Segmentation	10.1

4.5 Discussion

One of the major advantages of the proposed heuristic approach is the reduced miss rate, particularly in the low threshold region. As expected, the miss rate generally decreases as the threshold decreases, as evident in Figure 4.6. However, it is interesting to note that this is not the case for the baseline system in the low threshold region. This behavour is a direct result of the single threshold approach used in the baseline segmentation algorithm. When the threshold decreases beyond the ideal range, missed boundary detections increase due to the value of

the threshold falling below the trough between the true peaks. This is illustrated in Figure 4.8, where the straight horizontal line represents the low segmentation threshold. The lower of the two true peaks therefore becomes a missed boundary. This is not an issue for the proposed heuristic algorithm, as the relative depth of the trough with respect to the height of the neighbouring peaks is used to determine whether speaker boundaries exist at each peak location. This demonstrates the robustness of the proposed algorithm compared to using a single threshold for segmentation.

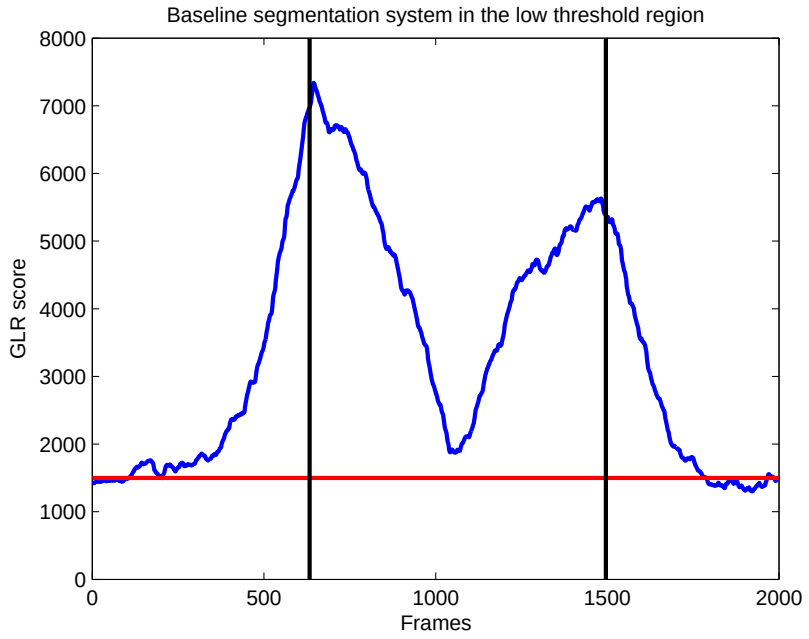

Figure 4.8: Segmentation based on a single low threshold: the lower peak on the right becomes a missed boundary

4.6 Summary

This chapter outlined a novel heuristic approach to speaker segmentation, with the emphasis of minimizing missed boundary detections. Due to the intrinsically noisy nature of the GLR distance curve resulting in false peaks with both high and low GLR scores, single threshold based segmentation is insufficient to ensure accurate segmentation. By analysing the relative depths of the troughs between

peaks rather than relying solely on the absolute GLR value of the peaks, the proposed heuristic approach was shown to be more robust compared to single threshold based segmentation. An algorithm for incorporating a smaller window and merging the two sets of speaker boundaries was also introduced to minimize missed boundary detections around short speaker segments.

Experiments conducted on the RT-02 Evaluation Dataset demonstrated improved segmentation performance across a wide range of thresholds. The miss rate was reduced across all thresholds, particularly in the low threshold region. However, the false alarm rate generally increased in the low threshold region. When tested in a diarization system, the proposed heuristic segmentation also led to an improvement in overall diarization performance compared to the baseline system.

The research presented in this chapter resulted in the following publication:

- **D. Wang**, R. Vogt, M. Mason and S. Sridharan, "Automatic audio segmentation using the generalized likelihood ratio," in *International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2008, pp. 1-5.

Chapter 5

Modelling Uncertainty in Speaker Model Estimates

5.1 Introduction

Speaker clustering, the process of associating segments of speech produced by the same speaker, is commonly performed as one of the final stages within a speaker diarization system. The clustering stage is responsible for associating all speech segments belonging to the same speaker by providing them with the same speaker label. Speaker clustering is commonly regarded as the most crucial step in the final stages of a speaker diarization system [74].

Recall from Section 2.8, one of the most popular speaker clustering strategies to date involves the use of a distance measure in conjunction with agglomerative clustering, otherwise known as bottom-up hierarchical clustering [74]. The choice of an appropriate distance measure is essential to the success of the clustering system using this approach. Various distance measures have been adopted for speaker clustering within diarization systems, including the popular Bayesian information criterion (BIC) [13], such as in [5, 52], and the Kullback-Leibler divergence [70], as in [68]. For hierarchical, agglomerative clustering approaches, BIC based clustering using single multivariate Gaussian modelling is arguably the most predominant approach [74], and is also used as the first clustering stage

in the baseline system described in Section 3.5.

Speaker clustering problems can be considered in the framework of statistical hypothesis testing. For two given speaker segments, the null hypothesis, H_0, is that the segments belong to different speakers, and the alternative hypothesis, H_1, is that they belong to the same speaker. Within this framework, the BIC can be seen as only an approximation to the Bayes factor of the marginal likelihoods of the data given each hypothesis [80]. The formulation of the BIC neglects the crucial prior term in the estimation of the marginal probabilities that make up the Bayes factor, thus foregoing the ability to incorporate prior beliefs about the expected distribution of parameter values. There is hence no guarantee that the marginal probability calculated using the BIC approximation would be close to the 'true' value, calculated using a prior distribution that would be regarded as appropriate by an observer. This chapter presents a speaker clustering technique based on the Bayes factor itself, with the aim of improving clustering performance and consistency by removing the errors caused by the BIC approximation.

The concept of using the Bayes factor as a decision criterion for clustering can be further extended to the speaker segmentation task. The Generalized likelihood ratio (GLR), used in the baseline system and throughout Chapter 4, is a likelihood ratio test statistic used to measure how much a segment boundary is favoured at one particular point of the audio. Previously used by Gish [31, 30] for the speaker identification task, the GLR is known as the most powerful likelihood ratio test, provided that the complete statistics are available. It is now one of the most popular distance metrics reported in speaker segmentation literature, such as in [15]. However, the use of the GLR involves the direct estimation of model parameters based on the maximum likelihood criterion, and does not take into account the uncertainty of the parameter estimates in the incomplete-data case. On the other hand, the Bayes factor of marginal likelihoods can be used as a distance metric which explicitly takes the uncertainty of the parameter estimates into account, based on the evidence of observations that did occur. This chapter thus extends the concept of using the Bayes factor as the distance measure to

the speaker segmentation task, applying the same heuristic approach detailed in Chapter 4 to determine which local maxima on the Bayes factor distance curve correspond to true speaker change points.

Section 5.2 reviews the popular maximum likelihood based criteria reported in speaker diarization literature, namely the GLR and the BIC, for speaker segmentation and clustering respectively. Section 5.3 presents the speaker clustering problem as a statistical hypothesis test, proceeding to develop the decision criterion for clustering under a Bayesian framework. This development results in the Bayes factor. Section 5.3.2 provides a detailed derivation of the BIC criterion, and shows the approximations made in the process in relation to the Bayes factor. The shortfalls of the BIC approximation are also outlined. The full solution to the marginal probability integral that make up the Bayes factor is then derived in Section 5.3.3, using multivariate Gaussian modelling. Section 5.3.4 outlines the method of estimating the hyperparameters associated with the Bayes factor expression, and Section 5.3.5 extends the concept of using the Bayes factor as a decision criterion to the speaker segmentation task.

Section 5.4 details the experiments performed and results achieved when comparing the traditional BIC based speaker clustering system to the proposed Bayes factor system. Extending the concept of using the Bayes factor as a decision criterion for speaker segmentation, results obtained using Bayes factor based segmentation is also compared to GLR based segmentation. Overall diarization results achieved using Bayes factor based segmentation and clustering is also presented and compared to the baseline system.

5.2 Review of Existing Distance Measures for Speaker Segmentation and Clustering

This section reviews two of the most commonly used distance measures reported in speaker diarization literature, GLR and BIC, for the segmentation and clustering tasks respectively.

5.2.1 The Generalized Likelihood Ratio

In the hypothesis testing framework, the GLR is formulated as a likelihood ratio test statistic which directly compares the likelihood of the two competing hypotheses. Let the data to be modelled be given by $\boldsymbol{X}_0 = \{x_i : i = 1, \cdots, N\}$, which is made up of two speaker segments, $\boldsymbol{X}_1 = \{x_i : i = 1, \cdots, m\}$ and $\boldsymbol{X}_2 = \{x_i : i = m+1, \cdots, N\}$. Let $\hat{M}_0$, $\hat{M}_1$ and $\hat{M}_2$ denote the models for the speaker segments, whose parameters are given by the maximum likelihood estimates calculated using the data $\boldsymbol{X}_0$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively. The GLR can be used to determine whether a speaker segment boundary exists between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In this case, the GLR is given by

$$GLR = \log \frac{p(\boldsymbol{X}_1|\hat{M}_1)p(\boldsymbol{X}_2|\hat{M}_2)}{p(\boldsymbol{X}_0|\hat{M}_0)}\,. \tag{5.1}$$

The denominator represents the likelihood of the combined data $\boldsymbol{X}_0$ given the combined model $\hat{M}_0$, and the numerator represents their separate counterparts. The GLR value determines how much a segment boundary is favoured between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The larger the GLR value, the more evidence that the two segments are more appropriately modelled by two separate distributions and should be segmented, and vice versa.

In the case of single multivariate Gaussian modelling, which will be used throughout this chapter, the GLR becomes

$$GLR = \log \frac{p(\boldsymbol{X}_1|\hat{\boldsymbol{\mu}}_{X_1}, \hat{\boldsymbol{\Sigma}}_{X_1})p(\boldsymbol{X}_2|\hat{\boldsymbol{\mu}}_{X_2}, \hat{\boldsymbol{\Sigma}}_{X_2})}{p(\boldsymbol{X}_0|\hat{\boldsymbol{\mu}}_{X_0}, \hat{\boldsymbol{\Sigma}}_{X_0})}\,. \tag{5.2}$$

Maximum likelihood estimation assumes that there is a 'correct' value for each model parameter. The correct value is the one that maximises the likelihood of the data. Due to the maximum likelihood parameter estimation, the GLR is the most powerful likelihood ratio test, provided that the complete statistics are available. Under each hypothesis, the distribution of the data is assumed to be fully specified.

5.2.2 The Bayesian Information Criterion

The concept behind using the BIC for speaker segmentation and clustering is very similar to that of the GLR. In the GLR expression, each of the likelihoods, $p(\boldsymbol{X}|M)$, is given by the maximum likelihood estimate for each model. The BIC estimate of each likelihood is mathematically equivalent to the maximum likelihood estimates used in constructing the GLR, except for the fact that the likelihoods are penalized by the model complexities, ie. the number of parameters used in the models. The BIC estimate for a given likelihood is therefore simply the maximum likelihood estimate for the model, minus a penalty term, which is a function of the number of model parameters. As a model selection criterion, the BIC is used to select an optimal model out of a number of candidate parametric models, which best represents a given data set. Since the maximum likelihood principle invariably leads to choosing a model with the highest possible dimension, the penalty term is used to penalize candidate models according to their complexities [69]. For any given segment, the general form for the BIC value is given by

$$BIC = p(\boldsymbol{X}|\hat{M}) - \frac{\lambda}{2} \cdot k \cdot \log N \,, \tag{5.3}$$

where $p(\boldsymbol{X}|\hat{M})$ is the maximum likelihood of the data $\boldsymbol{X}$ given the model M, λ denotes the BIC penalty factor, k is the number of parameters in the model M, and N is the number of samples in the data. In deciding whether or not two speaker segments belong to the same speaker, the variation of the BIC value between the two competing hypotheses is therefore given by

$$\Delta BIC = GLR - \frac{\lambda}{2} \cdot \Delta k \cdot \log N \,, \tag{5.4}$$

where Δk is the difference of the number of parameters between the two hypotheses. In BIC theory, the value of the penalty factor, λ, should be equal to 1 [13]. According to [13], BIC has the advantage of not requiring any thresholding. Ideally, speaker clustering decisions should be made based on whether the ΔBIC value is greater or less than 0. However, this is only true if $\lambda = 1$ or if there is a

systematic way to find the optimal value of λ. In absence of this, λ is an implicit threshold built into the penalty term [1].

In the case of single multivariate Gaussian modelling, the BIC score between the speaker segments $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be written as [13]

$$\Delta BIC = N \log |\boldsymbol{\Sigma}_0| - m \log |\boldsymbol{\Sigma}_1| - (N-m) \log |\boldsymbol{\Sigma}_2| - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right) \log N \,, \tag{5.5}$$

where d denotes the dimensionality of the feature vector space.

Like the GLR criterion, the Bayesian information criterion is also a maximum likelihood criterion. It is therefore not strictly a *Bayesian* model selection criterion, in the sense that it does not require or take into account any prior information, and nowhere does a prior appear in the expression for the BIC.

5.3 The Bayes Factor as a Distance Metric

This section presents the Bayes factor as a decision criterion for speaker clustering, in the framework of statistical hypothesis testing. The Bayes factor is a ratio of marginal likelihoods of the data given each hypothesis. A derivation of the BIC is presented, which shows the BIC as an approximation to the Bayes marginal likelihood integral, and the shortfalls associated with the approximation are outlined. This is followed by a derivation of the exact solution to the marginal probability integral that make up the Bayes factor, using multivariate Gaussian modelling. Finally, the concept of using the Bayes factor as a decision criterion is extended to the speaker segmentation task.

5.3.1 The Bayes Factor of Marginal Likelihoods

To derive an expression for the Bayes factor as a decision criterion for speaker clustering in the framework of statistical hypothesis testing, let the null hypothesis, H_0, be that the two segments should be modelled by two separate Gaussian distributions (and hence should be kept separate), and the alternative hypothesis,

H_1, be that the two segments are more appropriately modelled by one Gaussian distribution (and hence should be clustered). Let the data to be modelled be given by $\boldsymbol{X} = \{x_i : i = 1, \cdots, N\}$. According to Bayesian decision theory, the criterion based on which the clustering should be made is given by

$$\frac{p(H_1|\boldsymbol{X})}{p(H_0|\boldsymbol{X})}\,. \tag{5.6}$$

Applying Bayes Theorem, the posterior probability of each hypothesis given the data can be written as

$$p(H|\boldsymbol{X}) = \frac{p(H)p(\boldsymbol{X}|H)}{p(\boldsymbol{X})}\,. \tag{5.7}$$

Since $p(\boldsymbol{X})$ is identical for both hypotheses, it will not affect the hypothesis testing and will cancel out under the decision criterion given in (5.6). Also, assuming equal prior probability for each hypothesis (ie. $p(H_0) = p(H_1) = \frac{1}{2}$), $p(H)$ will also cancel out. Under these assumptions, $p(H|\boldsymbol{X})$ is proportional to the likelihood of the data given each hypothesis, and the Bayes factor, B, defined as a ratio of the likelihood of the data given the two competing hypotheses, can be written as

$$B = \frac{p(\boldsymbol{X}|H_1)}{p(\boldsymbol{X}|H_0)}\,. \tag{5.8}$$

Let the first speaker segment contain data $\boldsymbol{X}_1$ and the second contain $\boldsymbol{X}_2$. Let the single multivariate Gaussian distribution that supports H_1 be M_0, and the separate Gaussian distributions that support H_0 be M_1 and M_2 respectively. The Bayes factor given in (5.8) can then be written as

$$B = \frac{p(\boldsymbol{X}|H_1)}{p(\boldsymbol{X}|H_0)} = \frac{p(\boldsymbol{X}_1, \boldsymbol{X}_2|M_0)}{p(\boldsymbol{X}_1|M_1)p(\boldsymbol{X}_2|M_2)}\,. \tag{5.9}$$

As evident from (5.9), the larger the Bayes factor, the more evidence that the two segments are more appropriately modelled by one multivariate normal distribution and should be clustered, and vice versa.

To evaluate the Bayes factor, one must first derive an expression for each

of the terms on the right hand side of (5.9). Let $\boldsymbol{\theta}$ be the parameter set of the model under consideration. Rather than assuming there is a correct value for each model parameter as in the maximum likelihood approach, the Bayesian approach considers all model parameters as random variables, each with their own probability distribution of values that the parameters could be expected to have. The probability that the data conform to a model M, can be given by the marginal probability integral

$$p(\boldsymbol{X}|M) = \int p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta}\,. \tag{5.10}$$

The marginal probability can be interpreted as the expected value of the likelihood of the data given the model. It is given by the likelihood of the data given each model parameter set, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$, weighted by the associated prior probability distribution assumed for the model parameters, $p(\boldsymbol{\theta}|M)$, and integrated over all possible values of the parameters.

5.3.2 The BIC approximation

The marginal probability integral shown in (5.10) is difficult to compute when M has a large number of parameters, for example when M is a high-order Gaussian mixture model (GMM), as it involves integrating over a large number of parameters, and difficulties arise in deciding which data belongs to which mixture component. In this case, the BIC can be used as a relatively simple approximation to the marginal probability integral, by using the Laplace approximation to derive an expression for $\log p(\boldsymbol{X}|M)$. The Laplace approximation aims to approximate a probability density function defined over a set of continuous variables by finding a Gaussian approximation centred on a mode of the distribution [8]. Using the Laplace approximation, (5.10) can be written as

$$\log p(\boldsymbol{X}|M) \approx \log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, M) + \log p(\hat{\boldsymbol{\theta}}|M) + \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}|\,, \tag{5.11}$$

where $\hat{\boldsymbol{\theta}}$ is the set of parameters that maximises the log likelihood of the data under model M, which is denoted by $\log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, M)$. $\boldsymbol{A}$ is the Hessian matrix of second derivatives of the negative log posterior [8],

$$\boldsymbol{A} = -\nabla\nabla \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{X})\,. \tag{5.12}$$

Since $\boldsymbol{A} \approx N\boldsymbol{J}$, where $\boldsymbol{J}$ is the expected information matrix for a single observation and N is the number of data points, $\log|\boldsymbol{A}| \approx k\log(N) + \log|\boldsymbol{J}|$, where k is the number of parameters in $\boldsymbol{\theta}$. Equation (5.11) can therefore be written as

$$\begin{aligned}\log p(\boldsymbol{X}|M) \approx &\log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, M) + \log p(\hat{\boldsymbol{\theta}}|M) \\ &+ \frac{k}{2}\log(2\pi) - \frac{k}{2}\log(N) - \frac{1}{2}\log|\boldsymbol{J}|\,.\end{aligned} \tag{5.13}$$

Ignoring the second, third and last term in (5.13) gives the BIC approximation for the marginal log likelihood of the data given model M [8],

$$\text{BIC} = \log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, M) - \frac{k}{2}\log(N)\,. \tag{5.14}$$

As the above mathematical derivation shows, the BIC is only an approximation to the marginal probability integral. Given a constant number of model parameters, the third term of (5.13) is simply a constant offset and can thus be ignored without affecting the model selection. The effect of ignoring the last term is difficult to generalize, as it depends on the nature of the data as well as the parameterization. Arguably the biggest shortfall of the BIC approximation lies in ignoring the second term, which takes into account the prior beliefs about the expected distribution of parameter values. Ignoring the prior term means that the ability to incorporate one's prior beliefs about the expected distribution of parameter values has been lost. It is important to note here that since $p(\boldsymbol{X}|M)$ is a function of the prior distribution of the parameters, as shown in (5.10), there is potentially an infinite number of possible values for $p(\boldsymbol{X}|M)$, all of which are mathematically correct. However, not all possible values of $p(\boldsymbol{X}|M)$ can be

considered reasonable for this application, depending on the appropriateness of the prior distribution being chosen. The lack of ability to specify an appropriate prior distribution through the use of the BIC approximation means that the prior is implicitly inferred, and built into the calculated BIC value. For any given problem, some prior distributions may imply a value of $p(\boldsymbol{X}|M)$ very close to the value obtained by the BIC approximation, while others may imply a value that is far from it. The inability to specify a prior distribution through the BIC approximation therefore provides no guarantee that the value of the marginal likelihood calculated will be close to the 'true' marginal likelihood, calculated from a prior distribution that an observer would regard as appropriate for the data being modelled [80].

5.3.3 The Exact Solution to the Marginal Probability Integral

When a complex speaker model is used, such as a high order GMM, the marginal probability integral shown in (5.10) is difficult to evaluate as it involves integrating over a large number of parameters. In this case, the BIC is a practical method of approximating the marginal likelihood, despite its shortfalls outlined in Section 5.3.2. However, in the case of single mixture multivariate Gaussian modelling, as used in the baseline system described in Section 3.5 and throughout this chapter, the exact solution of the marginal probability integral can be used to construct the Bayes factor as a decision criterion for speaker clustering, which allows prior information to be incorporated and the uncertainty involved in the estimation of model parameters to be taken into account. It is hypothesized in this work that computing the exact solution to the marginal probability integral will give the best possible estimate for $p(\boldsymbol{X}|M)$.

To determine the exact solution to the integral in (5.10), one must first choose an appropriate distribution for $p(\boldsymbol{\theta}|M)$ to reflect one's prior beliefs about the expected distribution of parameter values. Following from common practice in speaker recognition and for simplicity, this work will consider only the means of

the distributions as the parameters in evaluating the marginal likelihood integral. The prior chosen here is therefore the prior on the mean. The variances will be estimated from the data itself, but treated as a known constant in the integral.

Since there are many factors that influence the prior distribution, such as the nature of the data, the parameterisation and the channel characteristics, the prior distribution on the mean is theoretically a sum of a large number of random variables. According to the central limit theorem, the mean of a large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. The form of the prior distribution chosen in this work is therefore the same as the model for the data, a multivariate normal distribution, as in [67].

In the case of a multivariate normal distribution being chosen as the model for the data as well as the prior, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ and $p(\boldsymbol{\theta}|M)$ in the marginal probability integral given in (5.10) can be expressed as

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{r} (\boldsymbol{x}_i - \boldsymbol{m})\right) \tag{5.15}$$

and

$$p(\boldsymbol{\theta}|M) = \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu})\right) \tag{5.16}$$

respectively, where $\boldsymbol{m}$ and $\boldsymbol{r}$ are the mean vector and precision matrix of the data, $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are the mean vector and precision matrix of the prior and D is the dimensionality of the feature vector. Considering only the mean vector $\boldsymbol{m}$ as the variable of integration, (5.10) becomes

$$\begin{aligned} p(\boldsymbol{X}|M) = \int \prod_{i=1}^{N} \left[\frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{r} (\boldsymbol{x}_i - \boldsymbol{m})\right) \right] \\ \cdot \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu})\right) d\boldsymbol{m}\,. \end{aligned} \tag{5.17}$$

While there is currently no known closed form solution to the indefinite integral given in (5.17), the definite integral over the entire space (ie. from $-\infty$ to

$+\infty$) is known and can be derived with the assistance of an appropriate table of integrals, such as [32].

For the ease of evaluating the integral, given that $\boldsymbol{r}$ and $\boldsymbol{\tau}$ are full precision matrices, simultaneous diagonalization can be used to transform the feature vector space so that $\boldsymbol{\tau}$ is whitened and $\boldsymbol{r}$ is diagonalized simultaneously in this new space. Each dimension can then be treated independently when evaluating the integral. Simultaneous diagonalization is achieved by finding a transformation matrix $\boldsymbol{A}$, to transform the data such that $\boldsymbol{X}' = \boldsymbol{A}\boldsymbol{X}$, where ($'$) denotes that the variable is expressed in the new space. As a result of this transformation, $\boldsymbol{\tau}'$ is the identity matrix and $\boldsymbol{r}'$ is diagonal.

The expression for the exact solution to the integral is given by

$$p(\boldsymbol{X}|M) = \prod_{d=1}^{D} \sqrt{\frac{(r'_d)^N}{(2\pi\Lambda_d)^N(Nr'_d+1)}} \exp\left\{ \frac{-Nr'_d}{2(Nr'_d+1)} \left[\frac{1}{N}\left(\sum_{i=1}^{N} (x'_{id} - \mu'_d)^2 \right) + Nr'_d \left(\frac{1}{N}\sum_{i=1}^{N}\left((x'_{id})^2\right) - \frac{1}{N^2}\left(\sum_{i=1}^{N}(x'_{id})\right)^2 \right) \right] \right\}, \tag{5.18}$$

where $\Lambda_1, \ldots, \Lambda_D$ are the eigenvalues of the prior covariance matrix in the original space. Note that $\boldsymbol{\tau}$ does not appear in this solution, due to the fact that $\boldsymbol{\tau}$ is whitened, and therefore $\boldsymbol{\tau}'$ is an identity matrix in the transformed space. Appendix A provides the detailed step-by-step derivations leading up to (5.18), starting with the simple single dimensional Gaussian case. Using the results obtained by the single dimensional case, the derivations expand to the multivariate case, using diagonal covariance. Finally, using the results obtained previously, the solution for the multivariate case using full covariance Gaussians is derived, resulting in (5.18).

The whitening of $\boldsymbol{\tau}$ and the diagonalization of $\boldsymbol{r}$, as opposed to the contrary, is a deliberate choice. Although both approaches will result in exactly the same marginal probability, whitening $\boldsymbol{\tau}$ allows the whitening matrix to be calculated only once. Conversely, whitening $\boldsymbol{r}$ would result in the need to calculate the whitening matrix for every segment being clustered. Whitening $\boldsymbol{\tau}$ is therefore

much less computationally expensive, and the processing time required for clustering using this approach is comparable to the BIC clustering approach used in the baseline system. Mathematical derivations involving the whitening of $\boldsymbol{r}$ and diagonalization of $\boldsymbol{\tau}$ is also included in Appendix A for academic completeness.

5.3.4 Estimating the Hyperparameters

This work proposes that the prior mean and precision can be estimated from the data itself. The prior mean estimate is given by the sample mean of all speech regions within the audio for a given show, ie. $\boldsymbol{\mu} = \overline{\boldsymbol{X}}$. The prior precision estimate for each segment is given by the sample precision matrix calculated from all speech regions of the audio, scaled by the number of samples in the segment, ie. $\boldsymbol{\tau} = N\boldsymbol{\Sigma}_X^{-1}$, using the central limit theorem as a guide. It is hypothesized in this work that estimating the prior from the data itself will ensure a true and accurate representation of the data distribution.

The value of the Bayes factor can now be calculated by substituting (5.18) into each term on the right hand side of (5.9), using the appropriate data and the corresponding parameter values. The resultant Bayes factor can then be used directly as a decision criterion for speaker clustering.

5.3.5 The Bayes Factor as a Decision Criterion for Speaker Segmentation

The theory developed in this chapter can also be applied to the speaker segmentation task, to determine how much a segment boundary is favoured at a certain point in the audio. In the statistical hypothesis testing framework, the null hypothesis, H_0, in this case is that the two segments either side of the boundary belong to the same speaker, and therefore the audio do not need to be segmented at this point. The alternative hypothesis, H_1, is that the two segments belong to different speakers and therefore a segment boundary is appropriate. In this

application, the Bayes factor becomes

$$B = \frac{p(\boldsymbol{X}|H_1)}{p(\boldsymbol{X}|H_0)} = \frac{p(\boldsymbol{X}_1|M_1)p(\boldsymbol{X}_2|M_2)}{p(\boldsymbol{X}_1, \boldsymbol{X}_2|M_0)}\,, \tag{5.19}$$

where each term on the right hand side of (5.19) is given by the Bayes marginal likelihood intergral in (5.10), and the solution presented in (5.18) can be used to evaluate the Bayes factor using exactly the same approach as for speaker clustering. Using a pair of adjacent sliding windows across the whole audio, the value of the Bayes factor is calculated at every point, producing a Bayes factor distance curve. Heuristic rules presented in Chapter 4 are then used to determine which local maxima points on the Bayes factor distance curve correspond to true speaker change points.

The GLR, used in the baseline system and throughout Chapter 4, is known as the most powerful likelihood ratio test due to the maximum likelihood parameter estimation, provided that the complete statistics are available. Under each hypothesis, the distribution of the data is assumed to be fully specified, and there are no uncertainties associated with the estimation of the model parameters. However, in this case, the complete statistics are not available due to the fact that the amount of data available in each speaker segment is limited, and hence there are uncertainties associated with the estimation of model parameters. The direct estimation of model parameters means that the GLR does not take into account this uncertainty. On the other hand, the Bayes factor considers the model parameters as random variables, each with their own probability distributions. This allows the uncertainties associated with the estimation of model parameters to be taken into account.

5.4 Experiments

This section presents the segmentation and clustering results obtained using the Bayes factor distance metric developed in this chapter, and compares the results to the baseline system. The overall diarization results, using Bayes factor based

segmentation and clustering, are also presented. The baseline system used in this study is the diarization system described in Section 3.5, with the segmentation stage replaced by the heuristic approach described in Chapter 4.

5.4.1 Speaker Clustering Experiments

To evaluate the proposed clustering strategy, the baseline system, which uses BIC clustering as the first clustering stage, is compared to a similar system with the BIC clustering stage replaced by the proposed Bayes factor clustering approach. For each of the 6 shows within the Rich Transcription 2002 (RT-02) Evaluation Dataset, the intermediate results obtained at the end of the first clustering stage are compared directly, using the average frame-level cluster purity and cluster coverage metrics calculated across all clusters. Cluster purity and cluster coverage metrics are defined in Section 3.2. Further details of the RT-02 Evaluation Dataset can be found in Section 3.3.

Table 5.1 shows the cluster purity and coverage results for each show at the end of the first clustering stage, using BIC and Bayes factor based clustering respectively. The average result of the 6 shows is calculated based on a time weighted average of the amount of scorable time in each show. The initial segments for clustering were produced using the baseline system processed up until the BIC clustering stage. Before clustering, the average cluster purity and coverage values across the 6 shows are 97.0% and 45.5% respectively. The low cluster coverage value is expected here, since speaker utterances are dispersed throughout the audio as unclustered segments at this stage. The cluster purity should ideally be 100% at this stage, and the loss is due to a small number of missed boundaries in the speaker segmentation stage. The results presented in Table 5.1 are based on the optimal operating points for each system; it corresponds to the optimal stopping threshold for this clustering stage, emperically tuned on each system to produce the best possible diarization error rate (DER) on this dataset.

Examining the results of each show individually, it is evident that the Bayes factor system is able to perform more merges than the BIC system, without

Table 5.1: Clustering Results - BIC vs Bayes Factor

Show	Coverage (%)		Purity (%)		# Merges	
	BIC	BF	BIC	BF	BIC	BF
1	61.0	61.2	95.2	95.5	169	173
2	68.6	65.8	98.3	98.3	152	155
3	97.8	76.3	99.2	99.2	88	83
4	80.8	78.2	90.0	90.3	81	86
5	62.3	60.6	98.3	98.5	132	135
6	63.3	63.1	96.8	96.1	81	87
Avg/Total	**71.9**	**67.4**	**96.2**	**96.3**	**703**	**719**

reducing the cluster purity. This is a desirable attribute, as it suggests that the Bayes factor system is able to cluster the segments further, thus bringing the system to a more complete clustering state, without introducing further clustering errors. This is particularly important in the first clustering stage, as clustering errors are very costly in this early stage. Erroneously clustered segments will result in clusters containing utterances from more than one speaker, which in turn contaminates statistics used in subsequent clustering, deteriorating subsequent clustering performance.

It is interesting to note the large discrepancy in cluster coverage for Show 3 between the two systems. Being two different distance metrics, the BIC and the Bayes factor will invariably produce different scores between the same pair of clusters, thus affecting the order by which the agglomerative clustering process is performed. The cluster coverage at the end of the first clustering stage is therefore dependent not only on the optimal stopping threshold, but also the order by which the segments are clustered. It can be argued that a lower cluster coverage at the end of the first clustering stage does not necessarily indicate suboptimal clustering performance, since the low cluster coverage can be rectified in the subsequent clustering stage.

5.4.2 Speaker Segmentation Experiments

Similar to the evaluation of the proposed clustering strategy presented in Section 5.4.1, the proposed segmentation strategy is evaluated by replacing the GLR based segmentation in the baseline system with the Bayes factor distance metric developed in this chapter. Both systems use the same sliding-window approach with the heuristic techniques proposed in Chapter 4. For the direct evaluation of the segmentation results obtained at the end of the segmentation stage, miss and false alarm rates are used, which are defined in Section 3.2. Miss and false alarm rates scored using 1 and 2 second intervals are reported in this work.

Table 5.2 shows the miss and false alarm rates for GLR versus Bayes factor based segmentation, both before and after Viterbi resegmentation, scored using a 1 second interval. For each system, the operating points were chosen based on the optimal configuration that produces the best overall DER. The effect of Bayes factor based segmentation and clustering on the overall DER is discussed in Section 5.4.3.

At the end of the segmentation stages, the miss rate is ideally 0%, since missed boundaries cannot be recovered in the clustering stages. As with incorrect clustering decisions, missed boundaries cause contaminated statistics in the clustering stage, deteriorating subsequent clustering performance. On the other hand, some false alarms can be tolerated at this stage, since correct clustering decisions will remove false alarms by clustering the utterances produced by the same speaker back together. The optimal configurations for each system therefore oversegment the audio before passing the segments into the clustering stages. However, excessive oversegmentation will produce many short segments containing insufficient statistics, which hinders the ability of the clustering stage to provide accurate clustering decisions. The choice of an optimal segmentation threshold is hence a trade-off between the desire to reduce the miss rate at the expense of increased false alarms, and the need to avoid producing unnecessary short segments.

From Table 5.2, it can be observed that the Bayes factor segmentation results before Viterbi resegmentation did not outperform those produced by the

Table 5.2: Segmentation Results (%) - 1 Second Interval

System		Miss Rate	FA Rate	Total Segs
Before Resegmentation	GLR	15.8	42.8	278
	BF	17.4	56.7	363
After Resegmentation	GLR	12.6	40.6	278
	BF	11.6	53.8	363

Table 5.3: Segmentation Results (%) - 2 Second Interval

System		Miss Rate	FA Rate	Total Segs
Before Resegmentation	GLR	6.8	36.3	278
	BF	4.7	50.1	363
After Resegmentation	GLR	6.3	36.3	278
	BF	4.7	50.3	363

GLR. However, it is interesting to note that the Viterbi resegmentation stage provided a greater benefit for the Bayes factor system through the refinement of segment boundary locations. In the case of Bayes factor based segmentation, more segments were relocated to correct speaker change points, resulting in a larger decrease in the miss rate. The results presented in Table 5.3 shows that the raw segmentation results produced by the Bayes factor system have a considerably lower miss rate than the GLR system, when scored using a 2 second interval. An interpretation of these results is that the Bayes factor system is more capable of hypothesizing segment boundaries and place them within the vicinity of true boundary locations, even though the accuracy of the segment boundary locations do not always fall within 1 second of the true boundary.

From the results shown in Tables 5.2 and 5.3, it is evident that the two systems use different operating points, which were chosen based on the optimal configuration that produces the best overall DER for each system. The Bayes factor system generates more segments, thus producing a lower miss rate at the expense of an increased false alarm rate. In order to compare the two systems directly, the segmentation threshold of the GLR system was decreased so that the miss rate matches that of the Bayes factor system before Viterbi resegmentation, when scored using a 2 second interval. The results are shown in Table 5.4. For

a given miss rate, the Bayes factor system achieved a 7.1% relative reduction in the false alarm rate.

Table 5.4: Segmentation Results (%) - Same Operating Point

System	Miss Rate	FA Rate	Total Segs
GLR	4.7	53.9	393
BF	4.7	50.1	363

5.4.3 Speaker Diarization Experiments

The overall diarization performance of the Bayes factor based segmentation and clustering systems, evaluated using the DER measure, is summarized in Table 5.5. Each system shown in Table 5.5 is tuned independently for optimal diarization performance. The 'Local' results shown are obtained by using the optimal local stopping threshold for each show in the final clustering stage, whereas the 'Global' results are obtained by using the same optimal global threshold across all shows that produces the best average DER, in accordance with the National Institute of Standards and Technology (NIST) evaluation protocol [19]. The average DER reported is calculated based on a time weighted average according to the amount of scorable time in each show. Details of the DER evaluation metric can be found in Section 3.2.

Table 5.5: Comparison of Overall Diarization Error Rates (%)

System	Local	Global
GLR Segmentation + BIC Clustering (Baseline)	10.1	12.5
GLR Segmentation + BF Clustering	9.5	10.4
BF Segmentation + BF Clustering	9.2	10.9

Compared to the baseline system, the differences in the outcome of the Bayes factor based clustering stage resulted in a considerable improvement in the overall DER. The Bayes factor based clustering system achieved relative improvements of 5.9% and 16.8% compared to the baseline system, when evaluated using local

and global thresholds respectively. These results suggest that the cluster coverage measure obtained at the end of the first clustering stage, as shown in Table 5.1, does not necessarily provide a useful indication of overall diarization performance. Despite a decreased cluster coverage, which indicates increased dispersion of speaker utterances across different clusters, the Bayes factor system achieved an improved DER compared to the baseline. This indicates that the second clustering stage is able to recover the dispersion of speaker utterances by completing the clustering process. On the other hand, the ability to perform more merges in the first clustering stage, without decreasing the cluster purity, appears to have more influence on diarization performance.

The concept of using the Bayes factor as a decision criterion is extended to the speaker segmentation task, replacing the GLR distance metric used for segmentation in the baseline system and throughout Chapter 4. This creates a speaker diarization system that uses the Bayes factor metric for both segmentation and clustering. Compared to the system using GLR segmentation and Bayes factor clustering, this system achieved a further 3.2% relative improvement in DER, when evaluated using local stopping thresholds. However, when evaluated using global stopping thresholds, the overall diarization performance deteriorated. This is due to the increased dispersion of optimal stopping thresholds across different shows in the Bayes factor segmentation and clustering system.

5.5 Summary

This chapter reviewed the speaker clustering problem as a statistical hypothesis test and developed the Bayes factor as a decision criterion under a Bayesian framework. The concept of using the Bayes factor as a decision criterion was then extended to the speaker segmentation task. As opposed to the popular maximum likelihood based criteria reported in speaker diarization literature, namely the BIC and the GLR, the proposed Bayes factor approach is able to incorporate prior information and take into account the uncertainty associated with the estimation

of speaker model parameters.

Experiments conducted on the RT-02 Evaluation Dataset demonstrated generally improved performance using Bayes factor based segmentation and clustering. Compared to BIC clustering, the Bayes factor system was able to perform more merges, thus bringing the system to a more complete clustering state, without introducing further errors. Compared to GLR segmentation, the Bayes factor system achieved improved segmentation performance when scored using a 2 second interval. However, when scored using a 1 second interval, it did not outperform the GLR system. When tested within a speaker diarization framework, the Bayes factor clustering system outperformed its BIC counterpart in terms of overall diarization performance, when using both local and global stopping thresholds. On the other hand, the Bayes factor segmentation system outperformed the GLR system only when scored using local stopping thresholds.

The research presented in this chapter resulted in the following publications:

- **D. Wang**, R. Vogt and S. Sridharan, "Bayes factor based speaker segmentation for speaker diarization," in *Interspeech*, 2010, pp. 1405-1408.

- **D. Wang**, R. Vogt and S. Sridharan, "Bayes factor based speaker clustering for speaker diarization," in *International Conference on Information Sciences, Signal Processing and their Applications (ISSPA)*, 2010, pp. 61-64.

Chapter 6

Modelling Uncertainty in Eigenvoice Modelling of Speaker Segments

6.1 Introduction

The improvements in speaker clustering performance through the use of the Bayes Factor, as detailed in Chapter 5, suggest that the incorporation of some prior knowledge about the whole audio appears to be beneficial for the speaker clustering task. A popular decision criterion which incorporates knowledge of the whole audio is the Cross Likelihood Ratio (CLR), which elegantly combines the information present in both clusters of interest with knowledge of the show background model. The use of the CLR for speaker clustering using Gaussian mixture models (GMM) is widely reported in speaker diarization literature, such as in [5, 56, 71].

Recently, eigenvoice modelling of speaker segments using Joint Factor Analysis (JFA) techniques has become increasingly popular in speaker recognition literature [44]. Compared to traditional GMM based approaches, which can potentially suffer from the lack of data caused by short speaker segments resulting in poor quality models, eigenvoice modelling has the advantage of being able to adequately represent a speaker with limited enrolment data [42]. This is achieved

by taking advantage of the highly informative prior distribution contained in the speaker models, and using only the most prominent eigenvoices, which account for most of the speaker variability. This greatly reduces the dimensionality and hence the number of parameters that need to be estimated [42]. JFA also has the potential to achieve improved capture of differences in speaker characteristics, through explicit and independent modelling of speaker and channel variations. Published speaker diarization systems based on eigenvoice modelling techniques are described in Section 2.9.

This work proposes that it would be beneficial to capitalize on the advantages of eigenvoice modelling in a CLR framework for speaker clustering, by incorporating eigenvoice modelling techniques into the CLR criterion. Motivated by the Bayesian approach to speaker clustering introduced in Chapter 5, which allows the uncertainties of speaker model estimates to be taken into account, Bayesian methods will also be used in this chapter to estimate the conditional probabilities in computing the CLR, with the aim of effectively combine the eigenvoice-CLR framework with the advantages of a Bayesian approach to the diarization problem.

Section 6.2 reviews the theory behind eigenvoice modelling. Section 6.3 reviews the CLR criterion, followed by detailed derivations showing how eigenvoice modelling techniques can be integrated into the CLR framework. Section 6.3.2 outlines the non-Bayesian approach, where the speaker factors are explicitly estimated and used directly in the calculation of the conditional probabilities in the CLR expression. This is followed by a detailed derivation of the Bayesian approach in Section 6.3.3, where the conditional probabilities are evaluated using the marginal likelihood integral, as described in Chapter 5. This results in a *Bayes-CLR* criterion, which interestingly reverts back to the form of a log likelihood ratio, except for a multiplicative normalization constant. The importance of this normalization constant and its effect on clustering performance is discussed in detail in Section 6.6.

The system implementation details, including how factor analysis models were

trained, is outlined in Section 6.4. Section 6.5 details the experiments performed and diarization results achieved when comparing the traditional GMM based clustering system to the proposed eigenvoice modelling system.

6.2 Review of Eigenvoice Modelling of Speaker Segments

As in some traditional speaker clustering approaches, eigenvoice modelling techniques are based around the use of GMMs to model a speaker. Let C be the number of mixture components in the GMM, and F be the dimensionality of the feature vector. From common practice in speaker recognition, only the GMM means are adapted during training. A GMM can therefore be conveniently expressed as a $CF \times 1$ supervector, obtained by concatenating the mean vectors of each mixture component.

Recall from Section 2.5, the eigenvoice modelling approach assumes that speaker supervectors have a Gaussian distribution of the form

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}\,, \tag{6.1}$$

where $\boldsymbol{s}$ represents a given speaker segment model, and $\boldsymbol{m}$ is a speaker independent universal background model (UBM) mean supervector obtained by the concatenation of the UBM component mean vectors. $\boldsymbol{V}$ is a $CF \times R$ matrix containing R basis supervectors in the eigenspace, often referred to as eigenvoices. While $R \ll CF$, it is assumed that the most prominent R eigenvoices contained in $\boldsymbol{V}$ are capable of capturing most of the speaker variability. This greatly reduces the dimensionality and hence the number of parameters that need to be estimated, allowing adequate speaker segment models to be constructed from limited enrolment data. $\boldsymbol{y}$ is a $R \times 1$ vector of speaker factors, which characterizes a unique speaker by specifying the amount of variability in each direction, defined by each of the R eigenvectors contained in $\boldsymbol{V}$, that is specific to that speaker.

The speaker variability model is trained such that $\boldsymbol{y}$ follows a standard normal distribution [40]. Overall, each unique speaker is characterized by the speaker-dependent offset $\boldsymbol{Vy}$ from the speaker-independent UBM mean supervector $\boldsymbol{m}$.

While the most prominent R eigenvoices have been shown to capture most of the speaker variability, adding a residual term $\boldsymbol{Dz}$ to the speaker model has proven beneficial in speaker recognition literature. By providing additional modelling power through the introduction of extra model parameters, the residual term aims to model any residual speaker variations that the speaker factor term fails to take into account [79]. The expression for a given speaker segment model then becomes

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{Vy} + \boldsymbol{Dz}\,. \tag{6.2}$$

Channel factors will not be included in the modelling of speaker segments as this work focuses on diarization of single-channel broadcast news data.

6.3 Incorporating Eigenvoice Modelling in the Cross Likelihood Ratio Framework

This section describes how eigenvoice modelling techniques can be integrated into the CLR framework for speaker clustering. A brief review of the CLR criterion as a similarity measure is first presented. A detailed mathematical derivation showing how eigenvoice modelling can be integrated into the CLR framework is then presented for both Bayesian and non-Bayesian approaches.

6.3.1 The Cross Likelihood Ratio Criterion

The CLR between two clusters, containing data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ respectively, is given in [5] as

$$\text{CLR} = \frac{1}{n_i} \log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)} + \frac{1}{n_j} \log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)}\,, \tag{6.3}$$

where n_i and n_j are the number of frames in each cluster, $p(\boldsymbol{x}|M)$ denotes the likelihood of the acoustic frames $\boldsymbol{x}$ given model M, and M_B represents the show

background model. This symmetric similarity measure elegantly combines the information present in both clusters of interest with knowledge of the show background model.

In the CLR equation, $\frac{1}{n_i}$ and $\frac{1}{n_j}$ serve as normalization constants, in order to compensate for the different amounts of data present in the clusters of interest. If the speech segments present in the two clusters are produced by the same speaker, $p(\boldsymbol{x}_i|M_j)$ and $p(\boldsymbol{x}_j|M_i)$ should be large, resulting in a large CLR value. Therefore, the larger the CLR, the more evidence that the two clusters should be merged into a single cluster, and vice versa.

6.3.2 The Cross Likelihood Ratio Decision Criterion using Eigenvoice Modelling: the non-Bayesian Approach

In order to describe how eigenvoice modelling can be integrated into the CLR framework, it is useful to first define some notations. Let $\boldsymbol{\Sigma}$ be the covariance of the speaker independent UBM; a $CF \times CF$ block diagonal matrix consisting of diagonal blocks Σ_c $(c = 1, \ldots, C)$, where Σ_c is the $F \times F$ diagonal covariance matrix corresponding to the mixture component c.

Next, the zeroth, first and second order statistics of the speaker segment need to be defined [40]. For each mixture component c $(c = 1, \ldots, C)$, let N_c be the total number of observations that are accounted for by the given mixture component, and let

$$F_c = \sum_t (x_t - m_c) \tag{6.4}$$

$$S_c = \text{diag}\left(\sum_t (x_t - m_c)(x_t - m_c)^* \right), \tag{6.5}$$

where the sums extend over all observations x_t that are aligned with the given mixture component, and m_c is the cth block of $\boldsymbol{m}$.

Let $\boldsymbol{N}$ be the $CF \times CF$ block diagonal matrix consisting of diagonal blocks N_c $(c = 1, \ldots, C)$. Let $\boldsymbol{F}$ be the $CF \times 1$ vector obtained by concatenating F_c $(c = 1, \ldots, C)$. Similarly, let $\boldsymbol{S}$ be the $CF \times CF$ block diagonal matrix consisting

of diagonal blocks S_c ($c = 1, \ldots, C$).

In the case where only the speaker factor term $\boldsymbol{Vy}$ is used to model the speaker segments, and assuming that $\boldsymbol{F}$ is centred around the UBM mean, it can be shown [40] that the log likelihood of the acoustic frames $\boldsymbol{x}$, given model M (in this case, the speaker factors $\boldsymbol{y}$), can be written as

$$\begin{aligned}\log p(\boldsymbol{x}|M) = &\sum_{c=1}^{C}\left(N_c \log \frac{1}{(2\pi)^{\frac{F}{2}}|\Sigma_c|^{\frac{1}{2}}}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) \\ &+ \boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F} - \frac{1}{2}\boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}\,. \end{aligned} \tag{6.6}$$

This expression can be broken down into two parts. The first two terms are dependent only on the data present in the speaker segment, whereas the last two terms also depend on the speaker model. This expression is not very easy to evaluate in its current form. However, the first two terms conveniently cancel out under the CLR formulation, due to the fact that each ratio making up the CLR rely on the same data. Under the CLR criterion, $\log p(\boldsymbol{x}|M)$ can hence be conveniently implemented as

$$\log p(\boldsymbol{x}|M) = \boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F} - \frac{1}{2}\boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}\,. \tag{6.7}$$

In order to use this result to construct the CLR as a decision criterion for speaker clustering, one must substitute (6.7) into each relevant term on the right hand side of (6.3), using the relevant data and speaker models. The end result is shown in (6.8). The CLR between two clusters i and j can be written as

$$\begin{aligned}\mathrm{CLR} = &\frac{1}{n_i}\left[(\boldsymbol{y}_j - \boldsymbol{y}_B)^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_i - \frac{1}{2}\boldsymbol{y}_j^*\boldsymbol{V}^*\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}_j + \frac{1}{2}\boldsymbol{y}_B^*\boldsymbol{V}^*\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}_B\right] \\ &+\frac{1}{n_j}\left[(\boldsymbol{y}_i - \boldsymbol{y}_B)^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_j - \frac{1}{2}\boldsymbol{y}_i^*\boldsymbol{V}^*\boldsymbol{N}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}_i + \frac{1}{2}\boldsymbol{y}_B^*\boldsymbol{V}^*\boldsymbol{N}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}_B\right], \end{aligned} \tag{6.8}$$

where M_i and M_j in (6.3) are represented by the speaker factors $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, which are enrolled using the data in the clusters i and j respectively. $\boldsymbol{y}_B$ is the

background speaker factors, enrolled using all speech segments from the whole show. This expression can now be used directly as a decision criterion for speaker clustering.

When both the speaker factors $\boldsymbol{Vy}$ and the residual term $\boldsymbol{Dz}$ are used to model speaker segments, similar derivations to the one shown above can be applied, with $\boldsymbol{Vy}$ replaced by $\boldsymbol{Vy}+\boldsymbol{Dz}$. Once again, assuming $\boldsymbol{F}$ is centred around the UBM mean, the resultant non-Bayesian CLR is given by

$$\begin{aligned}\text{CLR} = \frac{1}{n_i}\Bigg[&\Big((\boldsymbol{y}_j-\boldsymbol{y}_B)^*\boldsymbol{V}^*+(\boldsymbol{z}_j-\boldsymbol{z}_B)^*\boldsymbol{D}^*\Big)\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_i\\ &-\frac{1}{2}(\boldsymbol{y}_j^*\boldsymbol{V}^*+\boldsymbol{z}_j^*\boldsymbol{D}^*)\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}\boldsymbol{y}_j+\boldsymbol{D}\boldsymbol{z}_j)\\ &+\frac{1}{2}(\boldsymbol{y}_B^*\boldsymbol{V}^*+\boldsymbol{z}_B^*\boldsymbol{D}^*)\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}\boldsymbol{y}_B+\boldsymbol{D}\boldsymbol{z}_B)\Bigg]\\ +\frac{1}{n_j}\Bigg[&\Big((\boldsymbol{y}_i-\boldsymbol{y}_B)^*\boldsymbol{V}^*+(\boldsymbol{z}_i-\boldsymbol{z}_B)^*\boldsymbol{D}^*\Big)\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_j\\ &-\frac{1}{2}(\boldsymbol{y}_i^*\boldsymbol{V}^*+\boldsymbol{z}_i^*\boldsymbol{D}^*)\boldsymbol{N}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}\boldsymbol{y}_i+\boldsymbol{D}\boldsymbol{z}_i)\\ &+\frac{1}{2}(\boldsymbol{y}_B^*\boldsymbol{V}^*+\boldsymbol{z}_B^*\boldsymbol{D}^*)\boldsymbol{N}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}\boldsymbol{y}_B+\boldsymbol{D}\boldsymbol{z}_B)\Bigg]. \end{aligned} \tag{6.9}$$

6.3.3 The Cross Likelihood Ratio Decision Criterion using Eigenvoice Modelling: the Bayesian Approach

This section presents a detailed mathematical derivation of the Bayesian CLR, using only the speaker factors to model speaker segments. The Bayesian CLR incorporating the residual term is difficult to derive and is outside the scope of this work.

Computing the CLR using the expression given in (6.8) involves the enrolment of the speaker factors $\boldsymbol{y}$, and does not take into account the uncertainty associated with this direct estimation of $\boldsymbol{y}$. In order to model the uncertainty explicitly, $p(\boldsymbol{x}|M)$ can be evaluated by integrating out $\boldsymbol{y}$ using the Bayes Marginal Likelihood integral. For $p(\boldsymbol{x}_i|M_j)$, the integral is given by

$$p(\boldsymbol{x}_i|M_j) = \int p(\boldsymbol{x}_i|M_j, \boldsymbol{y}_j)p(\boldsymbol{y}_j|M_j)d\boldsymbol{y}_j\,. \tag{6.10}$$

Since the model M_j is the speaker factors $\boldsymbol{y}_j$, which comes from the data contained in cluster j, namely $\boldsymbol{x}_j$, (6.10) can be written as

$$p(\boldsymbol{x}_i|M_j) = \int p(\boldsymbol{x}_i|\boldsymbol{y}_j)p(\boldsymbol{y}_j|\boldsymbol{x}_j)d\boldsymbol{y}_j = p(\boldsymbol{x}_i|\boldsymbol{x}_j)\,. \tag{6.11}$$

In (6.11), $p(\boldsymbol{x}_i|\boldsymbol{y}_j)$ is the likelihood of $\boldsymbol{x}_i$ given $\boldsymbol{y}_j$. The expression for this likelihood is given in (6.7). $p(\boldsymbol{y}_j|\boldsymbol{x}_j)$ is the prior distribution of $\boldsymbol{y}_j$, which in this case is actually the *posterior* distribution of $\boldsymbol{y}$ after observing data $\boldsymbol{x}_j$. It can be shown [40] that the posterior distribution $p(\boldsymbol{y}_j|\boldsymbol{x}_j)$ is Gaussian with mean $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{\Sigma}_p$, where

$$\boldsymbol{\mu}_p = \boldsymbol{L}_j^{-1}\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_j \tag{6.12}$$

$$\boldsymbol{\Sigma}_p = \boldsymbol{L}_j^{-1}\,, \tag{6.13}$$

where $\boldsymbol{L}_j$ is the posterior precision of $\boldsymbol{y}$ after observing data $\boldsymbol{x}_j$, given by

$$\boldsymbol{L}_j = \boldsymbol{I} + \boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_j\boldsymbol{V}\,. \tag{6.14}$$

In (6.14), $\boldsymbol{I}$ is the identity matrix, which represents the prior precision of $\boldsymbol{y}$ before observing data $\boldsymbol{x}_j$. The integral given in (6.11) can now be written as

$$p(\boldsymbol{x}_i|\boldsymbol{x}_j) = \int p(\boldsymbol{x}_i|\boldsymbol{y}_j)N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)d\boldsymbol{y}_j\,. \tag{6.15}$$

It can be shown [40] that this expression is equivalent to

$$\begin{aligned}\log p(\boldsymbol{x}_i|\boldsymbol{x}_j) = &\sum_{c=1}^{C}\left(N_c \log \frac{1}{(2\pi)^{\frac{F}{2}}|\Sigma_c|^{\frac{1}{2}}}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_j\right)\\ &+ \log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big)N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)d\boldsymbol{y}_j\,.\end{aligned} \tag{6.16}$$

The expression given in (6.16) can be considered as the general case of the Bayes marginal likelihood integral presented in Theorem 3 of [40]. In [40], a solution to this integral is derived for the specific case where the prior distribution of $\boldsymbol{y}$ is assumed to be standard normal ie. $N(\mathbf{0}, \boldsymbol{I})$. In the case of the integral given in (6.16), the prior distribution of $\boldsymbol{y}$ is $N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ are defined in (6.12) and (6.13) respectively. While there is currently no known closed form solution to the indefinite integral given in (6.16), the definite integral over the entire space (ie. from $-\infty$ to $+\infty$) is known and can be derived with the assistance of an appropriate table of integrals. A detailed derivation is presented in Appendix A. The solution to the integral, in terms of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$, is given by

$$\begin{aligned}\log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big) N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) d\boldsymbol{y}_j = \\ -\frac{1}{2}\log|\boldsymbol{\Sigma}_p| - \frac{1}{2}\log|\boldsymbol{V}^*\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1}| \\ +\frac{1}{2}(\boldsymbol{F}_i^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\mu}_p^*\boldsymbol{\Sigma}_p^{-1})(\boldsymbol{V}^*\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1})^{-1} \\ (\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p) - \frac{1}{2}\boldsymbol{\mu}_p^*\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p\,. \end{aligned} \tag{6.17}$$

Once again, the first two terms on the right hand side of (6.16) rely on the same data, and hence cancel out under the CLR formulation. The result given in (6.17) can therefore be conveniently implemented as $\log p(\boldsymbol{x}_i|M_j)$ in the CLR equation.

The calculation of CLR using (6.17) first requires the computation of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$. In practice, computational efficiency can be improved by substituting the relevant parts of (6.17) by the appropriate posterior precision matrices $\boldsymbol{L}_i$ and $\boldsymbol{L}_j$ using (6.12) and (6.13), while keeping in mind that $\boldsymbol{L}_i = \boldsymbol{I} + \boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_i\boldsymbol{V}$ [40]. This eliminates the need to explicitly calculate $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$. The result of this substitution is shown in (6.18), which provides a mathematically equivalent

and computationally efficient way of computing $\log p(\boldsymbol{x}_i|\boldsymbol{x}_j)$.

$$\begin{aligned}\log p(\boldsymbol{x}_i|\boldsymbol{x}_j) = &\frac{1}{2}\log|\boldsymbol{L}_j| - \frac{1}{2}\log|\boldsymbol{L}_i - \boldsymbol{I} + \boldsymbol{L}_j| \\ &+\frac{1}{2}(\boldsymbol{F}_i + \boldsymbol{F}_j)^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}(\boldsymbol{L}_i - \boldsymbol{I} + \boldsymbol{L}_j)^{-1}\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}(\boldsymbol{F}_i + \boldsymbol{F}_j) \\ &-\frac{1}{2}\boldsymbol{F}_j^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{L}_j^{-1}\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_j \end{aligned} \tag{6.18}$$

As discussed previously, one powerful advantage of the CLR metric lies in its ability to combine the information present in both clusters of interest with knowledge of the show background model. In order to incorporate knowledge of the show background model effectively, it is convenient to first calculate the mean of the show background model, and translate the entire space to centre the origin at the location of the background mean. The CLR can then be evaluated in the translated space. It is worth noting that since the show background mean is calculated using all speaker segments from the whole show, it is expected that the value of the show background mean would lie somewhere in between the speaker segment models. This means that the show background mean carries useful prior information on what the speaker segments should look like. In the translated space, the individual speaker segment models can be interpreted as the speaker dependent offset from the show background mean.

For the calculation of $\log p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ in the translated space, since $\boldsymbol{V}$, $\boldsymbol{L}$ and $\boldsymbol{\Sigma}$ are all invariant under translation, only $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ need to be centred around the background mean, via

$$\boldsymbol{F'} = \boldsymbol{F} - \boldsymbol{N}\boldsymbol{M}_{bg}\,, \tag{6.19}$$

where $\boldsymbol{F'}$ denotes the first order statistics after background mean translation, and $\boldsymbol{M}_{bg}$ denotes the background mean supervector. Equation (6.18) can then be used directly to calculate $\log p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ in the translated space, with $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ replaced by $\boldsymbol{F'}_i$ and $\boldsymbol{F'}_j$.

In the translated space, $\log p(\boldsymbol{x}_i|M_B)$ in the CLR equation, given in (6.3), conveniently becomes $\log p(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I})$. The solution of the Bayes Marginal Likelihood integral with a standard normal prior given in [40] can hence be used to

directly evaluate $\log p(\boldsymbol{x}_i|M_B)$ in the new space,

$$\begin{aligned}\log p(\boldsymbol{x}_i|M_B) &= \log p(\boldsymbol{x}_i|\boldsymbol{0},\boldsymbol{I}) \\ &= -\frac{1}{2}\log|\boldsymbol{L}_i| + \frac{1}{2}\boldsymbol{F'}_i^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{L}_i^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{F'}_i\,. \end{aligned} \tag{6.20}$$

It is interesting to note that $p(\boldsymbol{x}_i|\boldsymbol{0},\boldsymbol{I})$ can simply be expressed as $p(\boldsymbol{x}_i)$, due to the fundamental assumption made in eigenvoice modelling, that the prior is standard normal before any data has been observed. The first ratio of the CLR equation can hence be written as

$$\begin{aligned}\log\frac{p(\boldsymbol{x}_i|\boldsymbol{x}_j)}{p(\boldsymbol{x}_i|\boldsymbol{0},\boldsymbol{I})} &= \log\frac{p(\boldsymbol{x}_i|\boldsymbol{x}_j)}{p(\boldsymbol{x}_i)} \\ &= \log p(\boldsymbol{x}_i,\boldsymbol{x}_j) - \log p(\boldsymbol{x}_i) - \log p(\boldsymbol{x}_j)\,. \end{aligned} \tag{6.21}$$

Similarly, the second ratio of the CLR equation can also be written as

$$\begin{aligned}\log\frac{p(\boldsymbol{x}_j|\boldsymbol{x}_i)}{p(\boldsymbol{x}_j|\boldsymbol{0},\boldsymbol{I})} &= \log\frac{p(\boldsymbol{x}_j|\boldsymbol{x}_i)}{p(\boldsymbol{x}_j)} \\ &= \log p(\boldsymbol{x}_i,\boldsymbol{x}_j) - \log p(\boldsymbol{x}_i) - \log p(\boldsymbol{x}_j)\,. \end{aligned} \tag{6.22}$$

A very interesting result arises from (6.21) and (6.22); the two ratios in the CLR equation are mathematically equivalent in the translated space. The CLR can therefore be written as

$$\begin{aligned}\mathrm{CLR} &= \frac{n_i+n_j}{n_i n_j}\Big[\log p(\boldsymbol{x}_i,\boldsymbol{x}_j) - \log p(\boldsymbol{x}_i) - \log p(\boldsymbol{x}_j)\Big] \\ &= \frac{n_i+n_j}{n_i n_j}\Big[\frac{1}{2}\log|\boldsymbol{L}_i| + \frac{1}{2}\log|\boldsymbol{L}_j| - \frac{1}{2}\log|\boldsymbol{L}_i - \boldsymbol{I} + \boldsymbol{L}_j| \\ &\qquad -\frac{1}{2}\boldsymbol{F'}_i^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{L}_i^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{F'}_i - \frac{1}{2}\boldsymbol{F'}_j^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{L}_j^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{F'}_j \\ &\qquad +\frac{1}{2}(\boldsymbol{F'}_i+\boldsymbol{F'}_j)^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}(\boldsymbol{L}_i-\boldsymbol{I}+\boldsymbol{L}_j)^{-1}\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}(\boldsymbol{F'}_i+\boldsymbol{F'}_j)\Big]\,. \end{aligned} \tag{6.23}$$

Equation (6.23) can now be used directly as a decision criterion for speaker

clustering. It can be viewed as the Bayesian version of the CLR expression given in (6.8).

6.3.4 Interpretation of the Bayes CLR

The interesting results presented in (6.21) and (6.22) provide a more meaningful insight into the interpretation of the quantity which the ratios $\log \frac{p(x_i|M_j)}{p(x_i|M_B)}$ and $\log \frac{p(x_j|M_i)}{p(x_j|M_B)}$ represent. It is interesting to note that in the translated space, the Bayes CLR essentially reverts back to a log likelihood ratio (with the exception of the normalization factor $\frac{n_i+n_j}{n_i n_j}$), despite the fact that the general CLR expression given in (6.3) was not formulated based on this theoretically optimal criterion. The Bayes CLR can be interpreted as the ratio of the likelihood that the data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ came from the same speaker model, over the likelihood that they came from different speaker models. This determines whether the two clusters of interest are more appropriately modelled by a combined model or two separate models, and hence whether it is appropriate to merge these clusters.

The Bayes CLR can therefore be regarded as a likelihood ratio consisting of three parts. The first part is the $\frac{n_i+n_j}{n_i n_j}$ normalization constant, the second part is a ratio of the determinant of the posterior precision matrix $\boldsymbol{L}$, and the last part is a ratio of the likelihood terms. Note that since

$$\boldsymbol{L}_i - \boldsymbol{I} + \boldsymbol{L}_j = \boldsymbol{I} + \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_j \boldsymbol{V} + \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_i \boldsymbol{V} \,, \tag{6.24}$$

$(\boldsymbol{L}_i - \boldsymbol{I} + \boldsymbol{L}_j)$ can be interpreted as the posterior precision matrix in the combined case, after data from both clusters i and j have been observed. Once again, the larger the overall value of (6.23), the more evidence that the two clusters should be merged.

The $\frac{n_i+n_j}{n_i n_j}$ normalization constant is responsible for the compensation of different amounts of data present in clusters i and j. Section 6.6 provides some discussions relating to the importance of this normalization constant and its effect on diarization performance. The second part can be interpreted as the

Bayes uncertainty contributions, with each $\log |\boldsymbol{L}|$ term providing an offset to the corresponding likelihood term in the third part. The magnitude of this offset represents the degree of uncertainty associated with the estimation of the corresponding likelihood. At the beginning of the merging process, when the amount of data present in the clusters are limited, the Bayes uncertainty contributions are expected to be relatively significant in magnitude compared to the likelihood part. As the merging progresses and the amount of data in the clusters increase, the uncertainty of the likelihood estimates decrease while the likelihood itself increases. With a large amount of data in the clusters, the Bayes uncertainty contributions become insignificant and the clustering decision is based primarily on the likelihood of the data.

6.4 System Implementation

The theory developed in this paper was tested against the final clustering stage of the baseline system, which uses a traditional GMM based modelling approach with the CLR decision criterion. To ensure a fair comparison, the new systems are identical to the baseline system up until the final clustering stage. The baseline system used in this study is the diarization system described in Section 3.5, using Bayes factor based segmentation and clustering as described in Chapter 5. The heuristic algorithm presented in Chapter 4 is also used in the segmentation stage.

Three separate systems were implemented, two intermediate systems and a final system. The first intermediate system was implemented using eigenvoice modelling techniques to adapt the UBM means for each speaker segment. The adapted supervectors are then converted back to a GMM and the CLR is evaluated, as in the baseline system. In the second intermediate system, eigenvoice modelling of speaker segments was integrated into the CLR framework, with and without the residual term $\boldsymbol{Dz}$, using the non-Bayesian CLR expressions given in (6.8) and (6.9) respectively. In the final system, the Bayesian CLR given in (6.23) was used as a decision criterion for speaker clustering. The results

obtained by the final system will hence demonstrate the overall relative improvement achieved by integrating eigenvoice modelling techniques into the Bayesian CLR framework; whereas the results obtained by the two intermediate systems will indicate how much of that improvement can be attributed to the advantages of eigenvoice adaptation over traditional maximum *a posteriori* (MAP) adaptation of speaker models, and the benefits of incorporating eigenvoice modelling in the CLR framework respectively.

In the new systems, the 512-mixture speaker independent UBM was trained using a total of approximately 5.5 hours of speech data, randomly selected from the 1996 and 1997 HUB4 English Broadcast News Corpus, as well as the 1996 USC Marketplace Corpus. Details of these databases can be found in Section 3.3. $\boldsymbol{V}$ was trained using utterances from 1165 speakers from the same databases, each of whom have at least 60 seconds of total speech. The large amount of data used to train $\boldsymbol{V}$ ensures a strong, highly informative prior on what the speaker model should look like. To prevent any dominant speakers from being overrepresented, data from all speakers with more than 5000 seconds of total speech were truncated to 5000 seconds for training. 300 principal eigenvoices were used to capture the speaker variability. $\boldsymbol{D}$ was trained using utterances from 30 speakers, also from the same databases, each with approximately 60 seconds of speech. In order to maximise the potential of the residual term to model any speaker variations that the speaker factor term fails to take into account, a disjoint set of speakers was used to train $\boldsymbol{D}$.

In the first intermediate system, a show background model was first adapted from the speaker independent UBM, using all speech segments from the whole show. Each initial cluster was then enrolled independently from the same UBM, with and without the residual term $\boldsymbol{Dz}$ in the eigenvoice modelling. This results in a mean-adapted supervector for each initial cluster. The cluster models were then converted back to traditional GMMs. Since only the means are adapted, the variances and mixture weights of the mean adapted cluster models are the same as that of the UBM. Agglomerative speaker clustering was then performed

using the CLR criterion. As in the baseline system, $p(\boldsymbol{x}|M)$ was calculated using the alignment scores of the acoustic frames with the associated model. In each iteration of the agglomerative clustering process, the CLR was calculated for each pair of potential merge candidates, and the closest pair of clusters were merged. This process is repeated until no more suitable merge candidates can be found.

In the second intermediate system, eigenvoice modelling techniques were integrated into the CLR framework, using (6.8) and (6.9). The background speaker factors $\boldsymbol{y}_B$ was first enrolled using all speech segments from the whole show. $\boldsymbol{z}_B$ was estimated at the same time when the residual term was included in the eigenvoice modelling. Each initial cluster was then enrolled, the value of CLR calculated between each pair of clusters, and agglomerative clustering performed. Once a merge is performed at the end of each iteration, a new $\boldsymbol{y}$ (and $\boldsymbol{z}$) was enrolled for the combined cluster using the combined data from both merge candidates. This new cluster then becomes a merge candidate in future iterations.

The final system can be regarded as the Bayesian equivalent of the second intermediate system, using only the speaker factors in the modelling of speaker segments. Once again, the background model $\boldsymbol{y}_B$ was first enrolled using all speech segments from the whole show. The show background mean supervector $\boldsymbol{M}_{bg}$ was then calculated using $\boldsymbol{M}_{bg} = \boldsymbol{V}\boldsymbol{y}_B$, to ensure that the show background mean is also constrained to the speaker subspace. The first order statistics were then centred around the background mean via (6.19). Agglomerative clustering was then performed using (6.23) as the decision criterion.

6.5 Experiments

This section presents the diarization results of the CLR based clustering approach using eigenvoice models, as obtained on the National Institute of Standards and Technology (NIST) Rich Transcription 2002 (RT-02) Evaluation dataset, and compares the results to the baseline system. Results obtained with and without the residual term $\boldsymbol{D}\boldsymbol{z}$ in eigenvoice modelling will be reported for both intermedi-

ate systems. For the final system, only the results obtained without the residual term will be reported, as the former case is outside the scope of this work.

Table 6.1 shows the overall diarization results for each system, without the residual term $\boldsymbol{Dz}$ in the eigenvoice modelling. The 'Local' results shown are obtained by using the optimal local stopping threshold for each show in the final clustering stage, whereas the 'Global' results are obtained by using the same optimal global threshold across all shows that produces the best average DER, in accordance with the NIST evaluation protocol [19]. The average result of the 6 shows is calculated based on a time weighted average according to the amount of scorable time in each show.

As evident from Table 6.1, the 'Global' results show an overall relative improvement in DER of 15.0% between the baseline system and the final system. The results obtained by the intermediate systems can be used to determine how much of this overall improvement can be attributed to each successive stage. The 5.3% relative improvement achieved by the first intermediate system over the baseline can be attributed to the use of eigenvoice adaptation of speaker models compared to traditional MAP adaptation. Comparing the second intermediate system to the first intermediate system, a further 6.3% relative improvement in DER was achieved. This can be attributed to the integration of eigenvoice modelling techniques into the CLR framework. Finally, a further 4.1% relative improvement was achieved by the final system over the second intermediate system, by using a Bayesian CLR measure. This improvement demonstrates the credibilty of the Bayesian approach and its ability to explicitly take into account the uncertainty involved in the direct estimation of $\boldsymbol{y}$, by means of marginalization. It is interesting to note that each successive system not only achieved a lower average DER compared to the previous system when evaluated using global thresholds, but the disparity between results obtained using local and global thresholds was also reduced. This is due to the optimal stopping thresholds being more similar across all shows in the latter systems. The robustness of the global stopping threshold provides more confidence that, when the system is

tuned on a development dataset and tested on a different evaluation dataset, a large drop in performance is less likely to occur.

Table 6.1: Diarization Error Rates (%) - No Residual Term

Show	Baseline		Intermediate1		Intermediate2		Final	
	Local	Global	Local	Global	Local	Global	Local	Global
1	14.15	14.15	9.94	11.38	7.54	7.54	9.09	9.09
2	7.35	11.32	6.51	6.51	6.19	6.51	6.19	6.19
3	0.70	0.75	0.33	2.95	0.97	0.97	1.19	1.19
4	7.41	12.67	10.68	10.68	11.62	11.62	16.21	16.21
5	9.15	9.74	3.74	4.54	3.93	4.80	4.42	5.30
6	15.29	15.29	23.04	25.12	25.24	25.59	16.68	16.68
Avg DER	**9.20**	**10.89**	**9.20**	**10.31**	**9.40**	**9.66**	**9.11**	**9.26**

Table 6.2 shows the overall diarization results for both intermediate systems, including the residual term $\boldsymbol{Dz}$ in the speaker segment model expression. The results obtained by both intermediate systems outperformed their counterparts shown in Table 6.1. This suggests that the additional modelling power introduced by incorporating the residual term in the modelling of speaker segments is beneficial for this application. When evaluated using global stopping thresholds, the first intermediate system achieved a relative improvement of 7.8% in DER over the baseline, while the second intermediate system achieved a further improvement of 13.6% over the first intermediate system. Overall, the second intermediate system achieved a 20.4% relative improvement in DER compared to the baseline system. Once again, each successive system not only achieved a lower average DER compared to the previous system, but the disparity between results obtained using local and global thresholds was also reduced.

6.6 Discussion

This section provides some discussions relating to the practical issues regarding the estimation of the show background mean, as well as the role of the $\frac{n_i+n_j}{n_i n_j}$ normalization constant in the CLR expression. As further analysis, the quality of the normalized Bayes CLR is directly compared to its unnormalized counterpart,

Table 6.2: Diarization Error Rates (%) - With Residual Term

Show	Baseline		Intermediate1		Intermediate2	
	Local	Global	Local	Global	Local	Global
1	14.15	14.15	10.02	10.02	7.83	11.38
2	7.35	11.32	6.51	6.51	6.19	6.51
3	0.70	0.75	0.34	0.67	0.33	0.33
4	7.41	12.67	9.45	11.62	10.68	10.68
5	9.15	9.74	3.93	4.80	3.74	3.74
6	15.29	15.29	20.87	25.59	18.46	18.46
Avg DER	**9.20**	**10.89**	**8.67**	**10.04**	**8.02**	**8.67**

namely the 'true' log likelihood ratio. All results presented in this section are compared to the 'Global' results achieved by the final system.

6.6.1 Background Mean Estimation

In the final system, the show background mean supervector was estimated in $\boldsymbol{V}$-space. The background model $\boldsymbol{y}_B$ was first enrolled, and the background mean supervector $\boldsymbol{M}_{bg}$ was calculated using $\boldsymbol{M}_{bg} = \boldsymbol{V}\boldsymbol{y}_B$. As an alternative, it is possible to estimate the background mean supervector using relevance MAP adaptation of the UBM, using all speech segments from the whole show, and concatenating the means of each mixture. Table 6.3 shows a comparison of results using the two approaches. The poorer result in the case of relevance MAP shows that it is important to constrain the show background mean to the speaker subspace, rather than allowing complete freedom through relevance MAP adaptation.

Table 6.3: Comparison of Background Mean Estimation Methods

System	Average Diarization Error Rate
$\boldsymbol{V}$-space	9.26
Relevance MAP	12.99

6.6.2 The Role of the Normalization Constant

From the general CLR expression given in (6.3), it is clear that the role of $\frac{1}{n_i}$ and $\frac{1}{n_j}$ is to normalize the number of frames in each of the individual clusters,

i and j. In the translated space, since the two ratios making up the CLR are mathematically equivalent, the normalization constant becomes $\frac{n_i+n_j}{n_i n_j}$, which is essentially a single constant multiplier applied to the log likelihood ratio, as given in (6.23). It is interesting to investigate the role of this normalization constant and examine how it affects overall diarization performance.

Table 6.4 shows a comparison of diarization performance between the final system and an identical system with the normalization constant removed from the Bayes CLR criterion. Interestingly, the system performance is sensitive to normalization, and keeping the normalization constant seems to be necessary to ensure optimal performance. Ideally, the criterion used for clustering should somehow penalize small clusters in order to favour the merge of two large clusters, since the merging of two large clusters is more likely to be correct. This is particularly important during the early stages of the clustering process, as erroneous merges cannot be rectified and the contaminated clusters subsequently pollute the statistics in future merges. In the Bayes CLR, the Bayes uncertainty terms (ie. the $\log|\boldsymbol{L}|$ terms) provide an offset to the value of the likelihood terms, thus penalizing the smaller clusters. However, it is evident from the results presented in Table 6.4 that without the normalization constant, the Bayes uncertainty terms themselves are unable to deliver optimal performance. One possible explanation for this observation is that, contrary to the fundamental assumption made in this system (and in fact, in most systems reported in speaker recognition literature to date), feature vectors are not strictly independent and identically distributed (iid) random variables. This is due to the overlap in the windows used in the feature extraction process, as well as the fact that the feature vectors themselves are dependent on the linguistic content of the data. Consequently, this causes the Bayes uncertainty terms to be overly confident. As the merging progresses and the size of the clusters increase, this over-confidence causes the Bayes uncertainty terms to become insignificant well before they ideally should be, as well as making the likelihood terms a greater magnitude than their fair value. This results in the lack of ability for the Bayes uncertainty terms to sufficiently penalize small

clusters without the aid of the normalization constant.

Table 6.4: Effect of Normalization on System Behaviour

System	Average Diarization Error Rate
Final	9.26
No Normalization	13.11
Scale $\boldsymbol{N}$&$\boldsymbol{F}$	12.31

One possible method of addressing the issue of over-confidence is to scale the zeroth and first order statistics ($\boldsymbol{N}$ and $\boldsymbol{F}$) by a factor of $\frac{1}{3}$, in order to compensate for the iid assumption regarding the feature vectors. A factor of $\frac{1}{3}$ was chosen since there is an overlap of approximately $\frac{2}{3}$ of the samples between consecutive feature vectors, due to the window size of 25 milliseconds and frame advance rate of 10 milliseconds used in the feature extraction process. It also produces the best results compared to other factors. This scaling of $\boldsymbol{N}$ and $\boldsymbol{F}$ could therefore theoretically eliminate the need to normalize the CLR by a factor of $\frac{n_i+n_j}{n_i n_j}$, by assuming that due to the devaluing of the weight of the likelihood terms relative to the Bayes uncertainty terms, the Bayes uncertainty terms alone would be able to sufficiently penalize the smaller clusters to achieve optimal performance. However, in practice this is not the case, as evident from the results presented in Table 6.4. Despite the slight improvement, the scaling of $\boldsymbol{N}$ and $\boldsymbol{F}$ is unable to significantly influence the behaviour of the clustering system, which seems to be far more sensitive to normalization. This is a practical issue; the removal of the normalization constant results in the smaller clusters being unfairly favoured. In practice, the combined value of all Bayes uncertainty terms is always positive, and the combined value of all likelihood terms is always negative. When the clusters are small, the Bayes uncertainty terms are significant compared to the likelihood terms, resulting in an overall CLR value closer to zero. When the clusters are large, the Bayes uncertainty terms become insignificant compared to the likelihood terms, and the overall CLR hence becomes a large negative value in the absence of appropriate cluster length normalization. This unfairly favours the smaller clusters, thus altering the system behaviour.

Interestingly, when one cluster is small and the other is large, they are still favoured over two large clusters in the absence of normalization. Let cluster i be the large cluster and cluster j be the small cluster. In this case, the value of $p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ is similar to $p(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I})$, due to the small amount of data in cluster j (ie. having access to little prior information through observing a small amount of prior data has very little effect compared to observing no prior data at all). This results in an overall CLR value close to zero. On the other hand, when both clusters are large, there is a significant difference between the values of $p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ and $p(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I})$. Once again, this results in a large negative value for the overall CLR in the absence of cluster length normalization, and therefore unfairly disfavouring the merging of the two large clusters. When cluster i is large and cluster j is small, it is interesting to note that the value of the normalization constant is approximately $\frac{1}{n_j}$. This means that the value of the normalization constant is dominated by the number of frames present in the smaller cluster. This is intuitive, since the amount of data present in cluster j determines how similar the value of $p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ is compared to $p(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I})$ (ie. how close $\boldsymbol{x}_j$ is to being non-informative). In the theoretical limit where the number of frames in cluster j approaches zero, the value of $p(\boldsymbol{x}_i|\boldsymbol{x}_j)$ approaches $p(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I})$.

In summary, the role of the normalization constant is to penalize small clusters, by means of individual length normalization. In practice, since the combined value of the likelihood terms are typically larger than the combined value of the Bayes uncertainty terms, the value of the overall CLR is generally negative. When clusters are small, the comparatively large value of $\frac{n_i+n_j}{n_i n_j}$ penalizes the small clusters and ensures that they are not unfairly favoured over larger merge candidates. Removing the normalization constant removes this penalty, and fundamentally alters the system behaviour by allowing small pairs of clusters to achieve overly favourable CLR scores. Since merges involving large clusters are more likely to be correct, large clusters should ideally be merged first to ensure optimal clustering performance. Despite the fact that the penalty introduced by the Bayes uncertainty terms should theoretically ensure this behaviour, in practice this penalty

disappears too early due to the over-confidence arising from the iid assumption made regarding the feature vectors. It is therefore proposed that the normalization constant need to be used in conjunction with the Bayes uncertainty terms for optimal clustering performance. This is also intuitive; in the sense that removing the normalization constant means the quantity being computed as a distance measure is no longer a valid CLR. The results presented in this section supports the notion that despite the two ratios in the CLR being the same in the transformed space, it does not warrant the removal of the normalization constant, which takes into account the number of frames associated with each cluster.

6.6.3 Comparison of Normalized and Unnormalized Bayes CLR Criteria

The discussions relating to the importance of the normalization constant so far have primarily focused on the penalty introduced by the normalization constant on smaller clusters, thereby altering the order of merges, and subsequently affecting clustering performance. This section provides some analysis aimed to directly compare the quality of the normalized and unnormalized criteria.

The quality of the criteria were compared using the set of underclustered nodes in the beginning of the second clustering stage. The value of the Bayes CLR, with and without normalization, was calculated for every possible pairwise combination within the scorable region of each show. For each criterion, the Bayes CLR values were ranked in descending order. A detection error trade-off (DET) plot was then produced to compare the quality of the criteria directly, using correct pairwise combinations as true scores and incorrect pairwise combinations as false scores. The DET plot shown in Figure 6.1 strongly favours the normalized criterion, as expected.

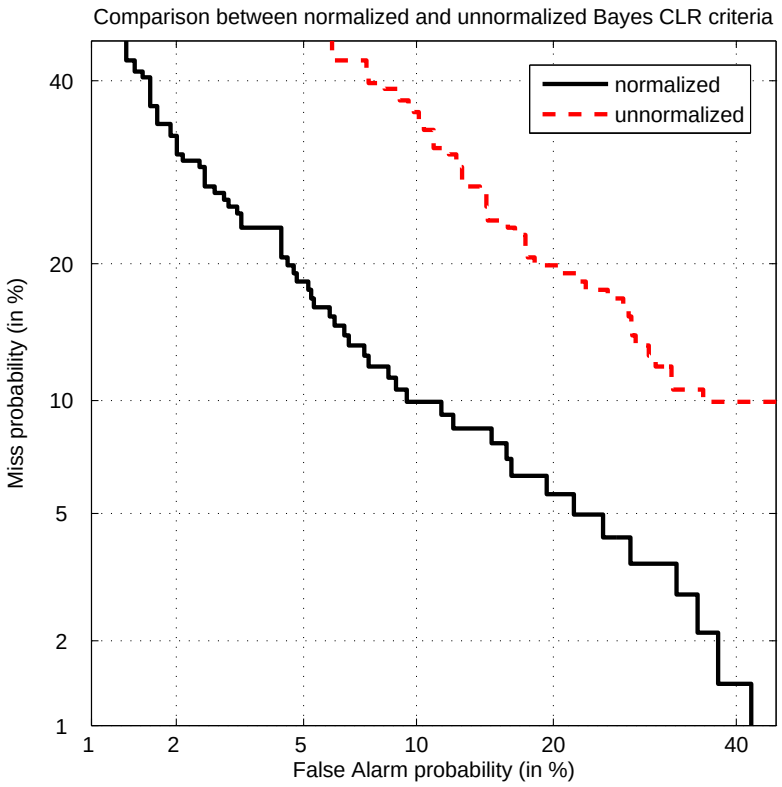

Figure 6.1: Comparison of normalized and unnormalized criteria

6.6.4 Further comments

All refinements to the Bayes CLR proposed in this section, including relevance MAP background mean estimation, removing the normalization constant and replacing the constant with a weighting factor to compensate for the high levels of correlation present in short-time acoustic features, have not resulted in improved diarization performance. However, it has been shown that the proposed Bayes CLR criterion not only outperforms the baseline as well as the non-Bayesian system, it also performs the speaker clustering task in a fully Bayesian manner. It can therefore be argued that the proposed Bayes CLR criterion can be interpreted as a theoretically optimal Bayesian criterion that requires no further refinements. This is supported by the results presented in this section, which shows that the attempts to refine this criterion appear to be unfavourable to diarization performance.

6.7 Summary

This chapter proposes the use of eigenvoice modelling techniques with the CLR criterion for speaker clustering within a speaker diarization system. By incorporating eigenvoice modelling into the CLR framework, it was possible to capitalize on the advantages of each technique to produce a robust speaker clustering system which outperforms traditional approaches using GMM based modelling. Bayesian methods have also been shown to be valuable in estimating the conditional probabilities in computing the CLR, thus effectively combining the eigenvoice-CLR framework with the advantages of a Bayesian approach to the diarization problem. Results obtained on the RT-02 Evaluation dataset show an improved clustering performance using the proposed approach, leading to a 15.0% relative improvement in the overall diarization performance compared to the baseline system. Furthermore, the results presented in Tables 6.1 and 6.2 suggest that including the residual term $\boldsymbol{Dz}$ in the modelling of speaker segments appears to be beneficial in this application. Through the use of intermediate systems, it was also possible to determine how much of the overall improvement can be attributed to the advantages of using eigenvoice modelling of speaker segments over traditional GMM based modelling approaches, the advantages of integrating eigenvoice modelling techniques into the CLR framework, and the benefits of estimating the CLR criterion using Bayesian methods.

The research presented in this chapter resulted in the following publications:

- **D. Wang**, R. Vogt and S. Sridharan, "Eigenvoice modeling for cross likelihood ratio based speaker clustering: a Bayesian approach," *Computer Speech and Language*, 2012. (Submitted)

- **D. Wang**, R. Vogt, S. Sridharan and D. Dean, "Cross likelihood ratio based speaker clustering using eigenvoice models," in *Interspeech*, 2011, pp. 957-960.

Chapter 7

Conclusion and Future Directions

7.1 Introduction

This chapter provides a summary of the work presented in this dissertation and the conclusions drawn. The summary follows the three main research themes and areas of contribution identified in Chapter 1 - heuristic rules for speaker segmentation, modelling uncertainty in speaker model estimates, and modelling uncertainty in eigenvoice speaker modelling.

7.2 Heuristic Rules for Speaker Segmentation

For distance metric based segmentation using a pair of sliding windows, the resultant distance curve is intrinsically noisy, and contains a large number of false peaks that do not correspond to true speaker boundaries. Since these false peaks can occur in regions with both high and low distance scores, the use of a single threshold is therefore insufficient to ensure accurate segmentation. Chapter 4 presented preliminary investigations into the development of a heuristic algorithm to determine which peaks on the distance curve correspond to true speaker change points, with the aim of minimizing missed boundary detections.

Original Contributions

- **A novel set of heuristic rules was proposed.**

  These heuristic rules govern the smoothing of the distance curve, detection of peak locations and the heuristic selection of 'significant' peaks from a list of candidate peaks. The decision to accept or reject a given peak as a true speaker boundary is based on analysing the relative depths of the trough between the peak itself and its neighbouring peak, as well as the absolute value of its distance score.

- **A second pass using a smaller pair of sliding windows was introduced for improved detection of boundaries around short speaker segments.**

  The second pass employs the same heuristic rules for peak detection and selection, and produces an additional set of segment boundary hypotheses. In combining the two sets of segment boundaries to produce the final set of speaker change points, a minimum duration constraint was imposed, which enforces a minimum number of frames between two speaker segments. This constraint ensures that the boundaries detected in the first pass are not detected again in the second pass as a different boundary located in close proximity to the initial boundary, providing some security that the speaker segments passed into the clustering stage will be of sufficient duration, in order to characterize the voice of the speakers reliably.

- **Experimental results demonstrated the effectiveness of the proposed heuristic algorithm for segment boundary decisions.**

  In contrast to a simple threshold based approach, the proposed heuristic algorithm provided improved segmentation performance across a wide range of thresholds. In particular, the miss rate was reduced across all thresholds. The improved segmentation performance led to a reduction in the overall diarization error rate.

7.3 Modelling Uncertainty in Speaker Model Estimates

If an infinite amount of speech data were available for the estimation of speaker models, it can be assumed that the estimated model parameters would be very close to their true values. However, this is never the case for practical diarization systems, as the amount of data in each speaker segment is always limited. This results in a degree of uncertainty associated with each model parameter estimate. Despite their popularity in speaker diarization literature, maximum likelihood based criteria do not take this uncertainty into account.

Chapter 5 reviewed the speaker segmentation and clustering processes as a statistical hypothesis test, and developed the Bayes factor as an optimal criterion under a Bayesian framework. The ability of the Bayesian approach to model the uncertainty of speaker model parameter estimates and to incorporate prior information to aid segmentation and clustering decisions was highlighted.

Original Contributions

- **A mathematical derivation of the popular Bayesian Information Criterion (BIC) was presented.**

  The BIC was presented as an approximation to the Bayes factor. The assumptions and approximations associated with the derivation were highlighted. Of particular interest was the omission of the prior term, which sacrifices its ability to incorporate prior information regarding the audio.

- **The Bayes factor, specific to multivariate, full covariance Gaussian speaker modelling, was developed as a distance metric for speaker clustering.**

  While it may be practical to resort to the BIC approximation when complex models such as high order GMMs are used to represent the speaker clusters, the Bayes factor can be constructed as a decision criterion in the case

of single multivariate Gaussian modelling, by solving the Bayes predictive density integral. The Bayes factor is suitable as a direct replacement for the BIC as a decision criterion for speaker clustering.

- **The concept of using the Bayes factor for speaker clustering was extended to the segmentation task.**

  The solution to the Bayes predictive density was also used to construct a decision criterion appropriate for segmentation. This criterion was applicable for use in place of the generalized likelihood ratio (GLR) for segmentation using the sliding-window approach. The heuristic algorithm presented in chapter 4 was also suitable in conjunction with the proposed Bayes factor metric.

- **Experiments demonstrated generally improved segmentation and clustering results using the proposed Bayes factor criterion.**

  This led to generally improved overall diarization performance.

Future Directions

The concept of using the Bayes factor can be extended to GMM based speaker modelling, to provide more detailed modelling of complex distributions. However, approximations are likely to be needed to overcome the difficulties arising as a result of the missing component occupancy information.

7.4 Modelling Uncertainty in Eigenvoice Speaker Modelling

Motivated by the success of joint factor analysis (JFA) based techniques for speaker modelling in the past few years, Chapter 6 investigated the idea of incorporating eigenvoice modelling techniques into the cross likelihood ratio (CLR) criterion in order to capitalize on the advantages of eigenvoice modelling in a CLR

framework for speaker clustering. Building on the work presented in Chapter 5, Bayesian methods were also developed to estimate the conditional probabilities in computing the CLR, which allows the uncertainties associated with the speaker factor estimates to be taken into account. This effectively combined the eigenvoice-CLR framework with the advantages of a Bayesian approach to the diarization problem.

Original Contributions

- **A novel, non-Bayesian criterion was developed, integrating eigenvoice modelling into the CLR framework. An alternative expression incorporating a residue term into the speaker model was also proposed.**

  The conditional probability distributions making up the CLR criterion are evaluated using their respective representations in the eigenvoice modelling framework, with and without the inclusion of the residue term. The proposed eigenvoice-CLR clustering approach was suitable for use in speaker clustering systems as a direct replacement for traditional CLR clustering with GMM based speaker modelling.

- **A Bayesian version of the eigenvoice-CLR criterion was also proposed, without the inclusion of the residue term in the modelling of speaker segments.**

  This allowed the uncertainty associated with the direct estimation of the speaker factors to be taken into account. The proposed Bayesian eigenvoice-CLR (referred to as Bayes CLR in the remainder of this chapter) criterion can be used in place of the eigenvoice-CLR criterion for speaker clustering. It was also discovered that after the translation of the space to centre the origin at the location of the background mean, the CLR essentially reverts back to a log likelihood ratio, with the exception of a multiplicative normalization constant.

- **Experiments demonstrated the effectiveness of the proposed metrics as decision criteria for speaker clustering.**

  Intermediate systems are also developed to indicate how much of the overall improvement can be attributed to each successive refinement stage leading up to the final Bayes CLR criterion. These include the replacement of traditional MAP adaptation of Gaussian mixture speaker models with eigenvoice adaptation, the integration of eigenvoice modelling techniques into the CLR framework, and employing a Bayesian approach for evaluating the conditional probabilities in the eigenvoice-CLR criterion.

- **Experimental results emphasized the importance of constraining the show background mean to the speaker subspace in the estimation process, rather than allowing complete freedom through relevance *maximum a posteriori* (MAP) adaptation.**

- **Experiments highlighted the importance of the normalization constant in the Bayes CLR criterion.**

  The normalization constant is responsible for penalizing small clusters by means of individual length normalization, which ensures that the small clusters are not unfairly favoured over larger merge candidates. Removing the normalization constant removes this penalty, and fundamentally alters the system behaviour by allowing small pairs of clusters to achieve overly favourable Bayes CLR scores.

- **A weighting factor was introduced to compensate for the high levels of correlation present in short-time acoustic features arising from the inherently untrue assumption of independently and identically distributed feature vectors.**

  The weighting factor, applied to the zeroth and first order statistics, was introduced with the aim of replacing the normalization constant in the Bayes CLR criterion in order to create a theoretically optimal log likelihood ratio

based metric. However, experimental results showed that despite a slight improvement in diarization performance compared to a non-normalized, non-weighted version of the Bayes CLR criterion, this weighting factor was unable to sufficiently influence the behaviour of the clustering system to fully compensate for the removal of the normalization constant. The system behaviour was discovered to be far more sensitive to normalization.

Future Directions

The results presented in Tables 6.1 and 6.2 suggest that including the residual term in the modelling of speaker segments appears to be beneficial for this application. It is therefore highly desirable to develop a Bayes CLR criterion including the residual term in the modelling of speaker segments to investigate whether further improvements in clustering performance can be achieved.

7.5 Summary

The overall aim of this work was to improve the performance and practicality of speaker diarization technology in the broadcast news audio domain through the reduction of diarization error rates. The segmentation and clustering stages within a diarization system were the focus of this research.

Three main research avenues were pursued: heuristic rules for speaker segmentation, modelling uncertainty in speaker model estimates, and modelling uncertainty in eigenvoice speaker modelling.

The novel research presented in this dissertation began with a preliminary investigation into the design of a heuristic approach for speaker segmentation, which governs the detection of candidate speaker segment boundary locations, as well as the selection of appropriate boundaries from the list of detected boundaries. The remainder of this research was largely dedicated to the development of novel decision criteria for speaker segmentation and clustering, based on evaluating the statistical similarity between two speaker segments. The proposed criteria, us-

ing Bayesian approaches, were shown to provide generally robust and improved segmentation and clustering performance compared to their non-Bayesian counterparts. This is true for classical GMM based speaker modelling approaches, as well as the more recently proposed JFA based approaches.

Improvements in segmentation and clustering performance introduced by the techniques developed in this research also led to improved overall diarization performance, offering increased potential for applications such as assisting automatic speech recognition and facilitating speaker indexing systems.

7.6 Future Work

The research presented in this dissertation focussed on the speaker diarization task, which is responsible for determining 'who spoke when' within a given audio recording. The relative speaker labels for each segment of speech is only relevant within the given recording, and do not pertain to any real-world speaker identities. Speaker diarization is therefore only the first step towards a fully functional data mining system that allows efficient searching, indexing and accessing of information from a large collection of spoken audio documents.

The next logical step towards the development of such a system therefore involves extending the existing speaker diarization technologies to the field of speaker attribution. As well as annotating segments of speech within an audio according to the speaker identities, a speaker attribution system also aims to associate speaker identities across different audio recordings. This relatively unexplored task can potentially provide a whole new level of structured information to spoken audio and enhance the user's ability to index, search and extract intelligent information from large spoken audio collections, resulting in improved accessibility for the search and retrieval of information contained in large audio archives. This provides another significant step towards determining the real-world identity of the detected speakers. The Bayesian framework developed in this research for speaker diarization can be extended to the speaker attribution

task.

As discussed in Chapter 1, universal speech unit models, trained on large databases with a large number of speakers, are used for speaker independent speech recognition in generic ASR systems. In this case, speaker attribution systems can be used to assist ASR systems by localizing the instances of specific speakers within and across different audio recordings to pool data for model adaptation, allowing speaker dependent speech recognition to be performed, which in turn boosts transcription accuracies. Conversely, text-dependent speaker recognition is also more accurate than text-independent speaker recognition. The speaker attribution and ASR tasks can therefore be considered co-dependent and mutually beneficial. To date, however, the interrelation between these two tasks has not been effectively exploited, but rather the extraction of spoken text and speaker identity annotation has been addressed as two distinct problems. There is hence significant potential to address the speaker attribution and ASR tasks as a joint problem under a unified Bayesian framework to realize the mutual benefits of both tasks, and further improve the automatic process of indexing, information retrieval and searching of large collections of spoken audio documents.

Bibliography

[1] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, 2004.

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2003, pp. 411–416.

[3] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Univ. Politecnica de Catalunya, Barcelona, Spain, 2006.

[4] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[5] C. Barras, Z. Xuan, S. Meignier, and J. Gauvian, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1505–1512, 2006.

[6] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *International Conference on Speech and Language Processing*, 2004.

[7] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and

D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[8] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Science and Business Media, LLC, 2006.

[9] S. Bozonnet, N. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 speaker diarization system: Enhancements in speaker modelling and cluster purification," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 4958–4961.

[10] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[11] A. Canavan and D. Miller, "USC Marketplace Broadcast News transcripts," Linguistic Data Consortium, Philadelphia. Available from http://www.ldc.upenn.edu, 1999.

[12] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2008, pp. 4133–4136.

[13] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] P. Delacourt and C. Wellekens, "DISTBIC : A speaker-based segmentation for audio data indexing," *Speech Communication*, pp. 111–126, 2000.

[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[17] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York City, New York, USA: John Wiley and Sons Inc, 2001.

[19] J. Fiscus, "Fall 2004 Rich Transcription RT-04F evaluation plan," National Institute of Standards and Technology, 2004.

[20] J. Fiscus, J. Garofolo, A. Le, A. Martin, D. Pallett, M. Przybocki, and G. Sanders, "Results of the Fall 2004 STT and MDE Evaluation," in *Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[21] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "1997 English Broadcast News speech (HUB4)," Linguistic Data Consortium, Philadelphia. Available from http://www.ldc.upenn.edu, 1998.

[22] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[23] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[24] R. Gangadharaiah, B. Narayanaswamy, and N. Balakrishnan, "A novel method for two speaker segmentation," in *International Conference on Speech and Language Processing*, 2004.

[25] J. Garofolo, J. Fiscus, and W. Fisher, "Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora," in *DARPA Speech Recognition Workshop*, 1997, pp. 15–21.

[26] J. Garofolo, J. Fiscus, and A. Le, "2002 Rich Transcription broadcast news and conversational telephone speech," Linguistic Data Consortium, Philadelphia. Available from http://www.ldc.upenn.edu, 2004.

[27] J. Garofolo, J. Fiscus, A. Martin, D. Pallett, and M. Przybocki, "NIST Rich Transcription 2002 evaluation: A preview," in *International Conference on Language Resources and Evaluation*, 2002, pp. 655–659.

[28] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *International Conference on Spoken Language Processing*, 1998, pp. 1335–1338.

[29] J.-L. Gauvain and C.-H. Lee, "Bayesian adaptive learning and MAP estimation of HMM," in *Automatic Speech and Speaker Recognition: Advanced Topics*, pp. 83-107, Boston, Massachusetts: Kluwer Academic, 1996.

[30] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 18–32, 1994.

[31] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 1991, pp. 873–876.

[32] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. San Diego: Academic Press, 2000.

[33] D. Graff, "An overview of broadcast news corpora," *Speech Communication*, vol. 37, pp. 15–26, 2002.

[34] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett, "1996 English Broadcast News speech (HUB4)," Linguistic Data Consortium, Philadelphia. Available from http://www.ldc.upenn.edu, 1997.

[35] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, "Segment generation and clustering in the HTK broadcast news transcription sys-

tem," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998. Available from http://mi.eng.cam.ac.uk/reports/abstracts/hain_darpa98.html.

[36] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities, second edition.* Cambridge, England: Cambridge University Press, 1988.

[37] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[38] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," in *International Conference on Speech and Language Processing*, 2004.

[39] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2000, pp. 1423–1426.

[40] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Computer Research Institute of Montreal. Available from http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf, Tech. Rep. CRIM-06/08-13, 2005.

[41] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," 2008. Available from http://www.crim.ca/perso/patrick.kenny.

[42] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Audio, Speech and Language Processing*, vol. 13, no. 3, pp. 345–359, 2005.

[43] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[44] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal on Selected Topics In Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[45] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[46] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, 1996.

[47] G. McLachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, NY, 1988.

[48] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 175–180.

[49] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, no. 20, pp. 303–330, 2006.

[50] N. Mirghafori and C. Wooters, "Nuts and flakes: a study of data characteristics in speaker diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2006, pp. 1017–1020.

[51] Y. Moh, P. Nguyen, and J.-C. Junqua, "Toward domain independent clustering," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, 2003, pp. 85–88.

[52] D. Moraru, M. Ben, and G. Gravier, "Experiments on speaker tracking and segmentation in radio broadcast news," in *European Conference on Speech Communication and Technology*, 2005.

[53] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation

during the NIST 2002 speaker recognition evaluation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2003, pp. 89–92.

[54] K. Mori and S. Nakagawa, "Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 413–416.

[55] P. Nguyen, L. Rigazio, Y. Moh, and J. Junqua, "Rich transcription 2002 site report. Panasonic speech technology laboratory (PSTL)," in *2002 Rich Transcription Workshop (RT-02)*, Available from http://www.nist.gov/speech/tests/rt/rt2002/presentations/rt02.pdf.

[56] M. Nishida and T. Kawahara, "Speaker model selection based on the bayesian information criterion applied to unsupervised speaker indexing," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 583–592, 2005.

[57] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1993.

[58] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[59] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[60] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, vol. 2, 1997, pp. 963–966.

[61] D. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. 4072–4075.

[62] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Interspeech*, 2009, pp. 1047–1050.

[63] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[64] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[65] D. Reynolds, E. Singer, B. Carlson, G. OLeary, J. McLaughlin, and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *International Conference on Spoken Language Processing*, 1998, pp. 3193–3196.

[66] D. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[67] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[68] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in *ICASSP*, vol. 5, 2006, pp. 521–524.

[69] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[70] M. Siegler, U. Jian, B. Rag, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[71] R. Sinha, S. Tranter, M. Gales, and P. Woodland, "The Cambridge University March 2005 speaker diarization system," in *Interspeech*, 2005, pp. 2437–2440.

[72] S. Stevens and J. Volkman, "The relation of pitch to frequency: A revised scale," *American Journal of Psychology*, vol. 52, pp. 329–353, 1940.

[73] S. Tranter, M. Gales, R. Sinha, S. Umesh, and P. Woodland, "The development of the Cambridge University RT-04 diarization system," in *Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[74] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[75] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Eurospeech'99*, vol. 2, 1999, pp. 679–682.

[76] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Eurecom, Sophia-Antipolis, France, 2005.

[77] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 260–269, 1967.

[78] R. Vogt, "Automatic speaker verification under adverse conditions," Ph.D. dissertation, Queensland University of Technology, Queensland, Australia, 2006.

[79] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, 2008, pp. 853–856.

[80] D. Weakliem, "A critique of the bayesian information criterion for model selection," *Socialogical Methods and Research*, vol. 27, pp. 359–397, 1999.

[81] J. Wilpon, L. Rabiner, and T. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 3, pp. 479–498, 1984.

[82] P. Woodland, "The development of the HTK Broadcast News transcription system: An overview," *Speech Communication*, vol. 37, pp. 47–67, 2002.

[83] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: the ICSI-SRI fall 2004 diarization system," in *Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[84] B. Zhou and J. Hansen, "Efficient audio stream segmentation via the combined T^2 Statistic and Bayesian information criterion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 467–474, 2005.

[85] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage speaker diarization for conference and lecture meetings," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

Appendix A

Supplemental Mathematical Derivations

A.1 Bayes Predictive Density Derivations

This section presents the step-by-step derivation of the solution to the Bayes marginal likelihood integral presented in Section 5.3.3, and arrives at the result given in Equation (5.18). The derivations begin with a simplified version of this problem, using single dimensional Gaussians, as presented in Section A.1.1. Section A.1.2 then expands the derivations to the multivariate case, using diagonal covariances. Using the results obtained previously, Section A.1.3 derives the solution for the multivariate case using full covariance Gaussians, and arrives at the result presented in Equation (5.18). As explained in Section 5.3.3, the result presented in (5.18) is derived using simultaneous diagonalization, by whitening the prior precision matrix and diagonalizing the data precision. In the clustering process, this deliberate choice allows the whitening matrix to be calculated only once. Conversely, despite the same end result in terms of the marginal likelihood, whitening the data precision matrix and diagonalizing the prior precision would result in the need to calculate the whitening matrix for every segment being clustered. The derivations for this scenario is presented in Section A.1.4. Due to the heavy computational demands associated with this approach, the solution

derived in Section A.1.4 is of little practical value for the application presented in this work, and is included here for academic completeness.

A.1.1 Single Dimensional Gaussians

Recall from Equation (5.10), the marginal probability integral is given by

$$p(\boldsymbol{X}|M) = \int p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta}\,. \tag{A.1}$$

In the single dimensional case, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ and $p(\boldsymbol{\theta}|M)$ can be expressed as

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - m)^2}{2\sigma^2}\right) \tag{A.2}$$

$$p(\boldsymbol{\theta}|M) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(\frac{-(m - \mu)^2}{2\sigma_p^2}\right), \tag{A.3}$$

where m and σ are the mean and standard deviation of the data, and μ and σ_p are the mean and standard deviation of the prior distribution. Given that only the means of the distributions will be considered, the marginal probability integral over the entire space (ie. from $-\infty$ to $+\infty$) can therefore be written as

$$p(\boldsymbol{X}|M) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - m)^2}{2\sigma^2}\right)\right] \cdot \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(\frac{-(m - \mu)^2}{2\sigma_p^2}\right) dm\,. \tag{A.4}$$

Taking all the terms that are not a function of m outside the integral and rearranging, step-by-step,

$$\begin{aligned} p(\boldsymbol{X}|M) &= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} \prod_{i=1}^{N} \left[\exp\left(\frac{-(x_i - m)^2}{2\sigma^2}\right)\right] \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(\frac{-(m - \mu)^2}{2\sigma_p^2}\right) dm \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N (2\pi)^{\frac{1}{2}}\sigma_p} \int_{-\infty}^{\infty} \exp\left[\sum_{i=1}^{N}\left(\frac{-(x_i - m)^2}{2\sigma^2}\right)\right] \exp\left(\frac{-(m - \mu)^2}{2\sigma_p^2}\right) dm \end{aligned}$$

$$
\begin{aligned}
p(\boldsymbol{X}|M) = & \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N(2\pi)^{\frac{1}{2}}\sigma_p}\int_{-\infty}^{\infty}\exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2}{2\sigma^2}\right)+\sum_{i=1}^{N}\left(\frac{mx_i}{\sigma^2}\right)-\frac{Nm^2}{2\sigma^2}\right]\cdot \\
& \exp\left(\frac{-m^2+2m\mu-\mu^2}{2\sigma_p^2}\right)dm \\
= & \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N(2\pi)^{\frac{1}{2}}\sigma_p}\int_{-\infty}^{\infty}\exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2}{2\sigma^2}\right)\right]\exp\left(\frac{m\sum_{i=1}^{N}x_i}{\sigma^2}\right) \\
& \exp\left(\frac{-Nm^2}{2\sigma^2}\right)\exp\left(\frac{-m^2}{2\sigma_p^2}\right)\exp\left(\frac{m\mu}{\sigma_p^2}\right)\exp\left(\frac{-\mu^2}{2\sigma_p^2}\right)dm \\
= & \frac{\exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2}{2\sigma^2}\right)\right]\exp\left(\frac{-\mu^2}{2\sigma_p^2}\right)}{(2\pi)^{\frac{N}{2}}\sigma^N(2\pi)^{\frac{1}{2}}\sigma_p}\cdot \\
& \int_{-\infty}^{\infty}\exp\left(-\frac{Nm^2}{2\sigma^2}-\frac{m^2}{2\sigma_p^2}+\frac{\sum_{i=1}^{N}(x_i)m}{\sigma^2}+\frac{\mu m}{\sigma_p^2}\right)dm \\
= & \frac{\exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2}{2\sigma^2}\right)\right]\exp\left(\frac{-\mu^2}{2\sigma_p^2}\right)}{(2\pi)^{\frac{N}{2}}\sigma^N(2\pi)^{\frac{1}{2}}\sigma_p}\cdot \\
& \int_{-\infty}^{\infty}\exp\left[-\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_p^2}\right)m^2+\left(\frac{\sum_{i=1}^{N}(x_i)}{\sigma^2}+\frac{\mu}{\sigma_p^2}\right)m\right]dm\,. \qquad \text{(A.5)}
\end{aligned}
$$

Using the identity

$$\int_{-\infty}^{\infty}\exp(-p^2x^2+qx)\,dx = \exp\left(\frac{q^2}{4p^2}\right)\cdot\frac{\sqrt{\pi}}{p}\,, \qquad \text{(A.6)}$$

and noting that

$$p^2 = \frac{N}{2\sigma^2}+\frac{1}{2\sigma_p^2} \qquad \text{(A.7)}$$

$$q = \frac{\sum_{i=1}^{N}(x_i)}{\sigma^2}+\frac{\mu}{\sigma_p^2}\,, \qquad \text{(A.8)}$$

the integral in (A.5) can be solved. (A.5) can therefore be written as

$$p(\boldsymbol{X}|M) = \frac{\exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2}{2\sigma^2}\right)\right]\exp\left(\frac{-\mu^2}{2\sigma_p^2}\right)}{(2\pi)^{\frac{N}{2}}\sigma^N(2\pi)^{\frac{1}{2}}\sigma_p}\cdot\exp\left[\frac{\left(\frac{\sum_{i=1}^{N}(x_i)}{\sigma^2}+\frac{\mu}{\sigma_p^2}\right)^2}{4\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_p^2}\right)}\right]\cdot\frac{\sqrt{\pi}}{\sqrt{\frac{N}{2\sigma^2}+\frac{1}{2\sigma_p^2}}} \qquad \text{(A.9)}$$

Simplifying and rearranging,

$$
\begin{aligned}
p(\boldsymbol{X}|M) = & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}(2\pi)^{\frac{1}{2}}} \exp\left[\sum_{i=1}^{N}\left(\frac{-x_i^2\sigma^{-2}}{2}\right)\right] \exp\left(\frac{-\mu^2\sigma_p^{-2}}{2}\right) \\
& \exp\left[\frac{\left(\sigma^{-2}\sum_{i=1}^{N}(x_i)+\mu\sigma_p^{-2}\right)^2}{2N\sigma^{-2}+2\sigma_p^{-2}}\right] \cdot \frac{\sqrt{\pi}}{\sqrt{\frac{N\sigma^{-2}+\sigma_p^{-2}}{2}}} \\
= & \frac{\sigma^{-N}\sigma_p^{-1}(2\pi)^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}(2\pi)^{\frac{1}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \\
& \exp\left[\frac{-\sigma^{-2}\sum_{i=1}^{N}(x_i^2)}{2} + \frac{-\mu^2\sigma_p^{-2}}{2} + \frac{\left(\sigma^{-2}\sum_{i=1}^{N}(x_i)+\mu\sigma_p^{-2}\right)^2}{2(N\sigma^{-2}+\sigma_p^{-2})}\right] \\
= & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\left[\frac{-1}{2(N\sigma^{-2}+\sigma_p^{-2})}\right. \\
& \left(\sigma^{-2}(N\sigma^{-2}+\sigma_p^{-2})\sum_{i=1}^{N}(x_i^2) + \mu^2\sigma_p^{-2}(N\sigma^{-2}+\sigma_p^{-2})\right. \\
& \left.\left. - \left(\sigma^{-2}\sum_{i=1}^{N}(x_i)+\mu\sigma_p^{-2}\right)^2\right)\right] \\
= & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\left[\frac{-N\sigma^{-2}}{2(N\sigma^{-2}+\sigma_p^{-2})}\right. \\
& \left(\frac{1}{N}(N\sigma^{-2}+\sigma_p^{-2})\sum_{i=1}^{N}(x_i^2) + \frac{\mu^2\sigma_p^{-2}}{N\sigma^{-2}}(N\sigma^{-2}+\sigma_p^{-2})\right. \\
& \left.\left. - \frac{1}{N\sigma^{-2}}\left(\sigma^{-2}\sum_{i=1}^{N}(x_i)+\mu\sigma_p^{-2}\right)^2\right)\right] \\
= & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\left[\frac{-N\sigma^{-2}}{2(N\sigma^{-2}+\sigma_p^{-2})}\right. \\
& \left(\left(\sigma^{-2}+\frac{\sigma_p^{-2}}{N}\right)\sum_{i=1}^{N}(x_i^2) + \mu^2\sigma_p^{-2} + \frac{\mu^2\sigma_p^{-4}}{N\sigma^{-2}}\right. \\
& \left.\left. - \frac{\sigma^{-2}}{N}\left(\sum_{i=1}^{N}(x_i)\right)^2 - \frac{2}{N}\sum_{i=1}^{N}(x_i)\mu\sigma_p^{-2} - \frac{\mu^2\sigma_p^{-4}}{N\sigma^{-2}}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{X}|M) = & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\Bigg[\frac{-N\sigma^{-2}}{2(N\sigma^{-2}+\sigma_p^{-2})} \\
& \left(\frac{\sigma_p^{-2}}{N}\sum_{i=1}^{N}(x_i^2) - \frac{2}{N}\sum_{i=1}^{N}(x_i)\mu\sigma_p^{-2} + \mu^2\sigma_p^{-2}\right. \\
& \left.+ \sigma^{-2}\sum_{i=1}^{N}(x_i^2) - \frac{\sigma^{-2}}{N}\left(\sum_{i=1}^{N}(x_i)\right)^2\right)\Bigg] \\
= & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\Bigg\{\frac{-N\sigma^{-2}}{2(N\sigma^{-2}+\sigma_p^{-2})} \\
& \left[\frac{\sigma_p^{-2}}{N}\left(\sum_{i=1}^{N}(x_i^2) - 2\sum_{i=1}^{N}(x_i)\mu + N\mu^2\right)\right. \\
& \left.+ N\sigma^{-2}\left(\frac{1}{N}\sum_{i=1}^{N}(x_i^2) - \frac{1}{N^2}\left(\sum_{i=1}^{N}(x_i)\right)^2\right)\right]\Bigg\} \\
= & \frac{\sigma^{-N}\sigma_p^{-1}}{(2\pi)^{\frac{N}{2}}\sqrt{N\sigma^{-2}+\sigma_p^{-2}}} \cdot \exp\Bigg\{\frac{-N\sigma^{-2}}{2(N\sigma^{-2}+\sigma_p^{-2})}\left[\frac{\sigma_p^{-2}}{N}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right)\right. \\
& \left.+ N\sigma^{-2}\left(\frac{1}{N}\sum_{i=1}^{N}(x_i^2) - \frac{1}{N^2}\left(\sum_{i=1}^{N}(x_i)\right)^2\right)\right]\Bigg\} \qquad \text{(A.10)}
\end{aligned}
$$

A.1.2 Multivariate Gaussians with Diagonal Covariance

In the multivariate case with D dimensions, the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ is given by

$$
p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N}\frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_i-\boldsymbol{m})\right), \qquad \text{(A.11)}
$$

where $\boldsymbol{m}$ and $\boldsymbol{r}$ are the mean vector and precision matrix of the data. Since $\boldsymbol{r}$ is diagonal, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ can be written as

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\theta}, M) = & \prod_{i=1}^{N}\frac{\prod_{d=1}^{D}(r_d)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}}\exp\left(-\frac{1}{2}\sum_{d=1}^{D}(x_{id}-m_d)^2 r_d\right) \\
= & \prod_{i=1}^{N}\frac{\prod_{d=1}^{D}(r_d)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}}\exp\left(-\sum_{d=1}^{D}\frac{r_d}{2}(x_{id}-m_d)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\theta}, M) &= \prod_{i=1}^{N} \frac{\prod_{d=1}^{D} (r_d)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \prod_{d=1}^{D} \exp\Big(-\frac{r_d}{2}(x_{id} - m_d)^2\Big) \\
&= \prod_{i=1}^{N} \prod_{d=1}^{D} \sqrt{\frac{r_d}{2\pi}} \exp\Big(-\frac{r_d}{2}(x_{id} - m_d)^2\Big) \,. \qquad \text{(A.12)}
\end{aligned}
$$

The prior $p(\boldsymbol{\theta}|M)$ in this case is given by

$$
p(\boldsymbol{\theta}|M) = \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\bigg(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu})\bigg) \,, \qquad \text{(A.13)}
$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are the mean vector and precision matrix of the prior. Since $\boldsymbol{\tau}$ is diagonal, this can be written as

$$
\begin{aligned}
p(\boldsymbol{\theta}|M) &= \frac{\prod_{d=1}^{D} (\tau_d)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\bigg(-\sum_{d=1}^{D} \frac{\tau_d}{2}(m_d - \mu_d)^2\bigg) \\
&= \prod_{d=1}^{D} \sqrt{\frac{\tau_d}{2\pi}} \exp\Big(-\frac{\tau_d}{2}(m_d - \mu_d)^2\Big) \,. \qquad \text{(A.14)}
\end{aligned}
$$

The marginal likelihood integral is therefore given by

$$
\begin{aligned}
p(\boldsymbol{X}|M) &= \int p(\boldsymbol{X}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) \, d\boldsymbol{\theta} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \Bigg[\prod_{i=1}^{N} \bigg(\sqrt{\frac{r_d}{2\pi}} \exp\Big(-\frac{r_d}{2}(x_{id} - m_d)^2\Big)\bigg) \cdot \\
&\quad \sqrt{\frac{\tau_d}{2\pi}} \exp\Big(-\frac{\tau_d}{2}(m_d - \mu_d)^2\Big) \Bigg] \, dm_1 \cdots dm_D \qquad \text{(A.15)}
\end{aligned}
$$

In the diagonal covariance case, all dimensions are independent. The results obtained from the single dimensional case, given in (A.10), can therefore be used directly to solve this integral, treating each dimension independently. The solu-

tion to the integral is given by

$$p(\boldsymbol{X}|M) = \prod_{d=1}^{D} \frac{\sqrt{r_d^N \tau_d}}{(2\pi)^{\frac{N}{2}}\sqrt{Nr_d + \tau_d}} \exp\left\{ \frac{-Nr_d}{2(Nr_d+\tau_d)} \left[\frac{\tau_d}{N}\left(\sum_{i=1}^{N} (x_{id} - \mu_d)^2 \right) + Nr_d \left(\frac{1}{N}\sum_{i=1}^{N}(x_{id}^2) - \frac{1}{N^2}\left(\sum_{i=1}^{N}(x_{id}) \right)^2 \right) \right] \right\} \tag{A.16}$$

A.1.3 Multivariate Gaussians with Full Covariance: Whitening the Prior

In the full covariance case, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ and $p(\boldsymbol{\theta}|M)$ can be expressed as

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^{\mathrm{T}} \boldsymbol{r} (\boldsymbol{x}_i - \boldsymbol{m}) \right) \tag{A.17}$$

$$p(\boldsymbol{\theta}|M) = \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu}) \right). \tag{A.18}$$

These look identical to the diagonal covariance case presented in Section A.1.2. However, in this case, $\boldsymbol{r}$ and $\boldsymbol{\tau}$ are no longer diagonal, which means that the dimensions of these multivariate Gaussians are not independent. In order to use the results derived so far to solve the marginal probablity integral directly, simultaneous diagonalization must be performed to transform the feature vector space such that $\boldsymbol{r}$ and $\boldsymbol{\tau}$ are both diagonal in the new space. This section presents the case where $\boldsymbol{\tau}$ is whitened and $\boldsymbol{r}$ is diagonalized simultaneously. Before the derivations are presented, it is useful to first review the simultaneous diagonalization process.

Simultaneous Diagonalization: Whitening Prior

Let $\boldsymbol{\Sigma}_p$ be the covariance of the prior distribution, and let $\boldsymbol{V}$ be a matrix of eigenvectors of $\boldsymbol{\Sigma}_p$. Let $\boldsymbol{D}$ be a diagonal matrix of the corresponding eigenvalues, such that $\boldsymbol{D} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}_p\boldsymbol{V}$.

The prior covariance, $\boldsymbol{\Sigma}_p$, is therefore written as

$$\boldsymbol{\Sigma_p} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}\,. \tag{A.19}$$

Define a whitening matrix, $\boldsymbol{W}$, such that $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$.

Let $\boldsymbol{B} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{W}$, where $\boldsymbol{\Sigma}$ is the covariance of the data, and let $\boldsymbol{Z}$ be a matrix of eigenvectors of $\boldsymbol{B}$, such that $\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{Z}$ is diagonal.

Now define the transformation matrix, $\boldsymbol{A}$, given by

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{Z}\,. \tag{A.20}$$

Under this transformation, $\boldsymbol{\tau}'$ is an identity matrix and $\boldsymbol{r}'$ is diagonal. Note that $(')$ denotes that the variable is expressed in the transformed space.

Solution to the Bayes Marginal Likelihood Integral

Transforming the data by $\boldsymbol{A}$ and compensating accordingly, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ can be written as

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{\theta}, M) &= \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{m})\right) \\ &= \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}'_i - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{x}'_i - \boldsymbol{m}')\right). \end{aligned} \tag{A.21}$$

Since $\boldsymbol{r}' = \boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}$, the data precision $\boldsymbol{r}$ is given by

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{r}'\boldsymbol{A}^{\mathrm{T}}\,. \tag{A.22}$$

Substituting $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{r}'\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{r}' = \boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}$ into (A.21) yields

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{(|\boldsymbol{A}||\boldsymbol{r}'||\boldsymbol{A}^{\mathrm{T}}|)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}'_i - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{r}'(\boldsymbol{x}'_i - \boldsymbol{m}')\right) \tag{A.23}$$

Since $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{Z}$, as given by (A.20), $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}$. Furthermore, since $\boldsymbol{W}^{\mathrm{T}} = \boldsymbol{W}$, $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$ and $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{Z}^{-1}$, it follows that $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{Z}^{-1}\boldsymbol{W}\boldsymbol{V}^{-1}$. (A.23) can therefore be written as

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{\theta}, M) =& \prod_{i=1}^{N} \frac{(|\boldsymbol{V}||\boldsymbol{W}||\boldsymbol{Z}||\boldsymbol{r}'||\boldsymbol{Z}^{-1}||\boldsymbol{W}||\boldsymbol{V}^{-1}|)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \\ & \exp\left(-\frac{1}{2}(\boldsymbol{x}_i' - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{r}'(\boldsymbol{x}_i' - \boldsymbol{m}')\right) \\ =& \prod_{i=1}^{N} \frac{|\boldsymbol{W}| \cdot |\boldsymbol{r}'|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i' - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{r}'(\boldsymbol{x}_i' - \boldsymbol{m}')\right) \end{aligned}$$

Noting that $\boldsymbol{r}'$ is diagonal, and that $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ can therefore be written as

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N}\prod_{d=1}^{D} \frac{1}{\sqrt{\Lambda_d}} \cdot \sqrt{\frac{r_d'}{2\pi}} \exp\left(-\frac{r_d'}{2}(x_{id}' - m_d')^2\right), \tag{A.24}$$

where $\Lambda_1, \ldots, \Lambda_D$ are the elements contained in $\boldsymbol{D}$, that is, the eigenvalues of the prior covariance matrix in the original space.

Similarly, transforming the prior mean by the transformation matrix $\boldsymbol{A}$ and compensating accordingly, $p(\boldsymbol{\theta}|M)$ can be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}|M) &= \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{m} - \boldsymbol{\mu})\right) \\ &= \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{m}' - \boldsymbol{\mu}')\right). \end{aligned} \tag{A.25}$$

Since the prior covariance is given by $\boldsymbol{\Sigma}_{\boldsymbol{p}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$, as given in (A.19), the prior precision can be written as

$$\boldsymbol{\tau} = (\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{-1}. \tag{A.26}$$

Noting that $\boldsymbol{A}^{-1} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{V}^{\mathrm{T}}$ and $(\boldsymbol{A}^{\mathrm{T}})^{-1} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{Z}$, substituting (A.26)

into (A.25) yields

$$p(\boldsymbol{\theta}|M) = \frac{(|\boldsymbol{V}||\boldsymbol{D}^{-1}||\boldsymbol{V}|^{-1})^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}} \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{W}^{-1} \boldsymbol{V}^{\mathrm{T}} \boldsymbol{V} \boldsymbol{D}^{-1} \boldsymbol{V}^{-1} \boldsymbol{V} \boldsymbol{W}^{-1} \boldsymbol{Z} (\boldsymbol{m}' - \boldsymbol{\mu}')\right). \tag{A.27}$$

Once again, noting that $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$, $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{Z}^{-1}$ and $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$, (A.27) can be simplified as

$$\begin{aligned} p(\boldsymbol{\theta}|M) =& \frac{|\boldsymbol{D}^{-1}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}} (\boldsymbol{m}' - \boldsymbol{\mu}')\right) \\ =& \prod_{d=1}^{D} \frac{1}{\sqrt{\Lambda_d}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(m'_d - \mu'_d)^2\right). \end{aligned} \tag{A.28}$$

Using the results obtained in (A.24) and (A.28), the marginal probability integral can be written as

$$\begin{aligned} p(\boldsymbol{X}|M) =& \int p(\boldsymbol{X}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta} \\ =& \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \left[\prod_{i=1}^{N} \left(\frac{1}{\sqrt{\Lambda_d}} \cdot \sqrt{\frac{r'_d}{2\pi}} \exp\left(-\frac{r'_d}{2}(x'_{id} - m'_d)^2\right)\right) \cdot \right. \\ & \left. \frac{1}{\sqrt{\Lambda_d}} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(m'_d - \mu'_d)^2\right)\right] dm_1 \ldots dm_D \\ =& \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \left[(\Lambda_d)^{-\frac{N}{2}} (\Lambda_d)^{-\frac{1}{2}} \prod_{i=1}^{N} \left(\sqrt{\frac{r'_d}{2\pi}} \exp\left(-\frac{r'_d}{2}(x'_{id} - m'_d)^2\right)\right) \cdot \right. \\ & \left. \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(m'_d - \mu'_d)^2\right)\right] dm_1 \ldots dm_D \,. \end{aligned} \tag{A.29}$$

The transformed mean, m'_d, cannot be integrated with respect to m_d. Performing the integration in (A.29) therefore requires a change of variables. Since $m'_d = \frac{1}{\sqrt{\Lambda_d}} m_d$, it follows that

$$dm'_d = \frac{1}{\sqrt{\Lambda_d}}\, dm_d \,. \tag{A.30}$$

The change of variables conveniently absorbs the $(\Lambda_d)^{-\frac{1}{2}}$ term in (A.29), giving

$$p(\boldsymbol{X}|M) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \left[(\Lambda_d)^{-\frac{N}{2}} \prod_{i=1}^{N} \left(\sqrt{\frac{r'_d}{2\pi}} \exp\left(-\frac{r'_d}{2}(x'_{id} - m'_d)^2 \right) \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(m'_d - \mu'_d)^2 \right) \right] dm'_1 \dots dm'_D \,. \tag{A.31}$$

All dimensions are independent in the transformed space. The solution to the diagonal covariance case, presented in Section A.1.2, can therefore be used to evaluate this integral, which yields the result presented in Equation (5.18),

$$p(\boldsymbol{X}|M) = \prod_{d=1}^{D} \sqrt{\frac{(r'_d)^N}{(2\pi\Lambda_d)^N (Nr'_d + 1)}} \exp\left\{ \frac{-Nr'_d}{2(Nr'_d + 1)} \left[\frac{1}{N} \left(\sum_{i=1}^{N} (x'_{id} - \mu'_d)^2 \right) + Nr'_d \left(\frac{1}{N} \sum_{i=1}^{N} \left((x'_{id})^2 \right) - \frac{1}{N^2} \left(\sum_{i=1}^{N} (x'_{id}) \right)^2 \right) \right] \right\} \tag{A.32}$$

as required. Note that the prior precision matrix does not appear in this expression, since it is the identity matrix in the transformed space.

A.1.4 Multivariate Gaussians with Full Covariance: Whitening the Data

This section presents the solution to the Bayes marginal likelihood integral, in the case where $\boldsymbol{r}$ is whitened and $\boldsymbol{\tau}$ is diagonalized.

Simultaneous Diagonalization: Whitening Data

Let $\boldsymbol{\Sigma}$ be the covariance of the data distribution, and let $\boldsymbol{V}$ be a matrix of eigenvectors of $\boldsymbol{\Sigma}$. Let $\boldsymbol{D}$ be a diagonal matrix of the corresponding eigenvalues, such that $\boldsymbol{D} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}$.

The data covariance, $\boldsymbol{\Sigma}$, is therefore written as

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1} \,. \tag{A.33}$$

Define a whitening matrix, $\boldsymbol{W}$, such that $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$.

Let $\boldsymbol{B} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma_p}\boldsymbol{V}\boldsymbol{W}$, where $\boldsymbol{\Sigma_p}$ is the covariance of the prior, and let $\boldsymbol{Z}$ be a matrix of eigenvectors of $\boldsymbol{B}$, such that $\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{Z}$ is diagonal.

Now define the transformation matrix, $\boldsymbol{A}$, given by

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{Z}\,. \tag{A.34}$$

Under this transformation, $\boldsymbol{r}'$ is an identity matrix and $\boldsymbol{\tau}'$ is diagonal. Note that $(')$ denotes that the variable is expressed in the transformed space.

Solution to the Bayes Marginal Likelihood Integral

Once again, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ is given by

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_i - \boldsymbol{m})\right). \tag{A.35}$$

Transforming the data by $\boldsymbol{A}$ and compensating accordingly, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ can be written as

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{\theta}, M) &= \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{m})\right) \\ &= \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}'_i - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{x}'_i - \boldsymbol{m}')\right). \end{aligned} \tag{A.36}$$

Since the data covariance is given by $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$, the data precision can be written as

$$\boldsymbol{r} = (\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{-1}\,. \tag{A.37}$$

Noting that $\boldsymbol{A}^{-1} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{V}^{\mathrm{T}}$ and $(\boldsymbol{A}^{\mathrm{T}})^{-1} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{Z}$, substituting (A.37)

into (A.36) yields

$$p(\boldsymbol{X}|\boldsymbol{\theta}, M) = \prod_{i=1}^{N} \frac{(|\boldsymbol{V}||\boldsymbol{D}^{-1}||\boldsymbol{V}^{-1}|)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i' - \boldsymbol{m}')^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{Z}(\boldsymbol{x}_i' - \boldsymbol{m}')\right). \tag{A.38}$$

Once again, noting that $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$, $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{Z}^{-1}$ and $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$, (A.38) can be simplified as

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{\theta}, M) &= \prod_{i=1}^{N} \frac{|\boldsymbol{D}^{-1}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i' - \boldsymbol{m}')^{\mathrm{T}}(\boldsymbol{x}_i' - \boldsymbol{m}')\right) \\ &= \prod_{i=1}^{N}\prod_{d=1}^{D} \frac{1}{\sqrt{\Lambda_d}}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_{id}' - m_d')^2\right), \end{aligned} \tag{A.39}$$

where $\Lambda_1, \ldots, \Lambda_D$ are the elements contained in $\boldsymbol{D}$, that is, the eigenvalues of the data covariance matrix in the original space.

The prior, on the other hand, is given by

$$p(\boldsymbol{\theta}|M) = \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\tau}(\boldsymbol{m} - \boldsymbol{\mu})\right). \tag{A.40}$$

Similarly, transforming the prior mean by the transformation matrix $\boldsymbol{A}$ and compensating accordingly, $p(\boldsymbol{\theta}|M)$ can be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}|M) &= \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{m} - \boldsymbol{\mu})\right) \\ &= \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{m}' - \boldsymbol{\mu}')\right). \end{aligned} \tag{A.41}$$

Since $\boldsymbol{\tau}' = \boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}$, the prior precision $\boldsymbol{\tau}$ is given by

$$\boldsymbol{\tau} = \boldsymbol{A}\boldsymbol{\tau}'\boldsymbol{A}^{\mathrm{T}}. \tag{A.42}$$

Substituting $\boldsymbol{\tau} = \boldsymbol{A}\boldsymbol{\tau}'\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{\tau}' = \boldsymbol{A}^{-1}\boldsymbol{\tau}(\boldsymbol{A}^{\mathrm{T}})^{-1}$ into (A.41) yields

$$p(\boldsymbol{\theta}|M) = \frac{(|\boldsymbol{A}||\boldsymbol{\tau}'||\boldsymbol{A}^{\mathrm{T}}|)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}}\boldsymbol{\tau}'(\boldsymbol{m}' - \boldsymbol{\mu}')\right) \tag{A.43}$$

Since $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{Z}$, as given by (A.34), $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}$. Furthermore, since $\boldsymbol{W}^{\mathrm{T}} = \boldsymbol{W}$, $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$ and $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{Z}^{-1}$, it follows that $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{Z}^{-1}\boldsymbol{W}\boldsymbol{V}^{-1}$. (A.43) can therefore be written as

$$\begin{aligned} p(\boldsymbol{\theta}|M) &= \frac{(|\boldsymbol{V}||\boldsymbol{W}||\boldsymbol{Z}||\boldsymbol{\tau}'||\boldsymbol{Z}^{-1}||\boldsymbol{W}||\boldsymbol{V}^{-1}|)^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}}\boldsymbol{\tau}'(\boldsymbol{m}' - \boldsymbol{\mu}')\right) \\ &= \frac{|\boldsymbol{W}| \cdot |\boldsymbol{\tau}'|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m}' - \boldsymbol{\mu}')^{\mathrm{T}}\boldsymbol{\tau}'(\boldsymbol{m}' - \boldsymbol{\mu}')\right) \end{aligned}$$

Noting that $\boldsymbol{\tau}'$ is diagonal, and that $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}$, $p(\boldsymbol{\theta}|M)$ can therefore be written as

$$p(\boldsymbol{\theta}|M) = \prod_{d=1}^{D} \frac{1}{\sqrt{\Lambda_d}} \cdot \sqrt{\frac{\tau'_d}{2\pi}} \exp\left(-\frac{\tau'_d}{2}(m'_d - \mu'_d)^2\right). \tag{A.44}$$

Using the results obtained in (A.39) and (A.44), the marginal probability integral can be written as

$$\begin{aligned} p(\boldsymbol{X}|M) &= \int p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \left[\prod_{i=1}^{N} \left(\frac{1}{\sqrt{\Lambda_d}} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x'_{id} - m'_d)^2\right)\right) \cdot \right. \\ &\quad \left. \frac{1}{\sqrt{\Lambda_d}} \cdot \sqrt{\frac{\tau'_d}{2\pi}} \exp\left(-\frac{\tau'_d}{2}(m'_d - \mu'_d)^2\right)\right] dm_1 \ldots dm_D \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \left[(\Lambda_d)^{-\frac{N}{2}} (\Lambda_d)^{-\frac{1}{2}} \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x'_{id} - m'_d)^2\right)\right) \cdot \right. \\ &\quad \left. \sqrt{\frac{\tau'_d}{2\pi}} \exp\left(-\frac{\tau'_d}{2}(m'_d - \mu'_d)^2\right)\right] dm_1 \ldots dm_D \,. \end{aligned} \tag{A.45}$$

The transformed mean, m'_d, cannot be integrated with respect to m_d. Performing the integration in (A.45) therefore requires a change of variables. Since $m'_d =$

$\frac{1}{\sqrt{\Lambda_d}}\, m_d$, it follows that

$$dm'_d = \frac{1}{\sqrt{\Lambda_d}}\, dm_d\,. \tag{A.46}$$

The change of variables conveniently absorbs the $(\Lambda_d)^{-\frac{1}{2}}$ term in (A.45), giving

$$p(\boldsymbol{X}|M) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{d=1}^{D} \Bigg[(\Lambda_d)^{-\frac{N}{2}} \prod_{i=1}^{N} \Bigg(\frac{1}{\sqrt{2\pi}} \exp\Bigg(-\frac{1}{2}(x'_{id} - m'_d)^2 \Bigg) \Bigg) \cdot \sqrt{\frac{\tau'_d}{2\pi}} \exp\Bigg(-\frac{\tau'_d}{2}(m'_d - \mu'_d)^2 \Bigg) \Bigg] dm'_1 \ldots dm'_D\,. \tag{A.47}$$

All dimensions are independent in the transformed space. The solution to the diagonal covariance case, presented in Section A.1.2, can therefore be used to evaluate this integral, giving

$$p(\boldsymbol{X}|M) = \prod_{d=1}^{D} \sqrt{\frac{\tau'_d}{(2\pi\Lambda_d)^N (N+\tau'_d)}} \exp\Bigg\{ \frac{-N}{2(N+\tau'_d)} \Bigg[\frac{\tau'_d}{N} \Bigg(\sum_{i=1}^{N} (x'_{id} - \mu'_d)^2 \Bigg) + N \Bigg(\frac{1}{N} \sum_{i=1}^{N} \Big((x'_{id})^2 \Big) - \frac{1}{N^2} \Big(\sum_{i=1}^{N} (x'_{id}) \Big)^2 \Bigg) \Bigg] \Bigg\}\,. \tag{A.48}$$

Note that the data precision matrix does not appear in this expression, since it is the identity matrix in the transformed space.

A.2 Bayes CLR Derivations

This section presents the step-by-step derivation of the solution to the Bayes marginal likelihood integral presented in Equation (6.14) in Section 6.3.3, and arrives at the result presented in Equation (6.15).

The integral is given in (6.14) as

$$\log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big) N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) d\boldsymbol{y}_j\,, \tag{A.49}$$

where the likelihood is given by

$$p(\boldsymbol{x}_i|\boldsymbol{y}_j) = \boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F} - \frac{1}{2}\boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y}\,, \tag{A.50}$$

and the prior is a k-dimensional Gaussian with mean $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{\Sigma}_p$, given by

$$N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}_p|^{\frac{1}{2}}} \exp\!\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_p)^T\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_p)\right). \tag{A.51}$$

Expanding the prior term and regrouping, (A.49) can be written as

$$\begin{aligned}\log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big) N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) d\boldsymbol{y}_j = \\ \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}_p|^{\frac{1}{2}}} \int \exp\Big(-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{V}^T\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1})\boldsymbol{y} \\ + \boldsymbol{y}^T(\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}\Big) d\boldsymbol{y}\,.\end{aligned} \tag{A.52}$$

Using the identity

$$\int \exp\Big(-\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{B} + \boldsymbol{C}\Big) d\boldsymbol{X} = \sqrt{\frac{\pi}{|\boldsymbol{A}|}} \exp\!\left(\frac{1}{4}\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{C}\right), \tag{A.53}$$

and noting that in this case

$$\boldsymbol{A} = \frac{1}{2}\big(\boldsymbol{V}^T\boldsymbol{N}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1}\big) \tag{A.54}$$

$$\boldsymbol{B} = \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu} \tag{A.55}$$

$$\boldsymbol{C} = -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}\,, \tag{A.56}$$

the $(2\pi)^{\frac{k}{2}}$ term conveniently cancels out, and the solution to the integral can be

written as

$$\log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big) N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) d\boldsymbol{y}_j = \frac{1}{|\boldsymbol{\Sigma}_p|^{\frac{1}{2}} |\boldsymbol{V}^T \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1}|^{\frac{1}{2}}} \exp\Bigg(\frac{1}{2} \big(\boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}\big)^T \big(\boldsymbol{V}^T \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1}\big)^{-1} \big(\boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}\big) - \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu} \Bigg). \quad \text{(A.57)}$$

Some simple algebraic manipulations then yield the result presented in Equation (6.15),

$$\begin{aligned} \log \int \exp\big(p(\boldsymbol{x}_i|\boldsymbol{y}_j)\big) N(\boldsymbol{y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) d\boldsymbol{y}_j = & \\ -\frac{1}{2} \log |\boldsymbol{\Sigma}_p| - \frac{1}{2} \log |\boldsymbol{V}^* \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1}| & \\ + \frac{1}{2} (\boldsymbol{F}_i^* \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \boldsymbol{\mu}_p^* \boldsymbol{\Sigma}_p^{-1}) (\boldsymbol{V}^* \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \boldsymbol{\Sigma}_p^{-1})^{-1} & \\ (\boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_i + \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}_p) - \frac{1}{2} \boldsymbol{\mu}_p^* \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}_p & \end{aligned} \quad \text{(A.58)}$$

as required.