http://researchspace.auckland.ac.nz

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage. http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Impact of the CDMA Mobile Phone Network on Speech Used for Forensic Voice Comparison

Esam A. S. Alzqhoul

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Faculty of Engineering

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Department of Electrical and Computer Engineering
The University of Auckland
2014
Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapters 3 and 4 of this thesis are based on our published work "Determination of Likelihood Ratios for Forensic Voice Comparison Using Principal Component Analysis" and this was published in the International Journal of Speech Language and the Law

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Extent of contribution by PhD candidate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refined the UKD algorithm for PCAKLR. Proposed the concept of PCAKLRNF, its implementation and its experimental evaluation. Wrote the sections that are related to PCAKLRNF in the Journal.</td>
<td>40</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balamurali Nair</td>
<td>Proposed the concept of PCAKLR, its implementation and its experimental evaluation. Wrote the sections that are related to PCAKLR in the Journal.</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td>Provided proper direction in setting up the methodology. Restructuring and proofreading the journal paper.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esam Alzqhou</td>
<td></td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Balamurali Nair</td>
<td></td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td></td>
<td>18/01/2015</td>
</tr>
</tbody>
</table>

Last updated: 25 March 2013
This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Section 5.1 and Appendix B of this thesis are based on the published work "An Alternative Approach for Investigating the Impact of Mobile Phone Technology on Speech", which was published in the International Conference on Signal Processing and Imaging Engineering (ICSPIE'14), San Francisco, USA, 2014

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Developed the CDMA codec platform and wrote its corresponding sections in the paper.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>40</td>
</tr>
</tbody>
</table>

### CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balamurali Nair</td>
<td>Developed the GSM codec platform and wrote its corresponding sections in the paper.</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td>Restructuring and proofreading the paper.</td>
</tr>
</tbody>
</table>

### Certification by Co-Authors

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esam Alzghoul</td>
<td>[Signature]</td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Balamurali Nair</td>
<td>[Signature]</td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td>[Signature]</td>
<td>18/01/2015</td>
</tr>
</tbody>
</table>

Last updated: 25 March 2013
Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 6 is based on our published paper "Comparison between Speech Parameters for Forensic Voice Comparison Using Mobile Phone Speech". This was published in the Speech Science and Technology (SST) Conference, Christchurch, New Zealand, 2014.

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Undertook the experiments related to the CDMA mobile phone recordings and wrote its corresponding sections.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>40</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balamurali Nair</td>
<td>Undertook the experiments related to the GSM mobile phone recordings and wrote its corresponding sections.</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td>Provided directions for setting up the methodology. Restructuring and proofreading the paper.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esam Alzqhoui</td>
<td></td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Balamurali Nair</td>
<td></td>
<td>18/01/2015</td>
</tr>
<tr>
<td>Bernard Guillemin</td>
<td></td>
<td>18/01/2015</td>
</tr>
</tbody>
</table>

Last updated: 25 March 2013
Abstract

This thesis investigates different aspects of the CDMA mobile phone network and quantifies their impact on the performance of Forensic Voice Comparison (FVC) analysis. The term FVC refers to the process of comparing suspect and offender voice samples in order to assess the strength of the speech evidence for a court of law. There exists an assumption among some forensic speech scientists that all mobile phone networks are similar with respect to their underlying technology, and therefore in their potential impact on the speech signal. However, within this arena there are a number of network providers utilizing a variety of technologies, such as the Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). These technologies are fundamentally different and they incorporate different mechanisms with respect to handling the speech signal. Therefore any assumption that they impact similarly on the speech signal is not correct. This thesis focuses on the CDMA network.

As will be discussed, the only component in these networks that can directly impact upon the quality of transmitted speech is the speech codec. Other factors such as poor wireless channel conditions, congestion related to the number of users, channel noise, etc., do not impact directly on the speech signal, but rather indirectly. In such cases, a set of instructions will be sent from the mobile phone network to the speech codec, which in turn will change its mode of operation to mitigate the impact of these factors.

With the above facts in mind, a new software platform has been developed as part of this research to simulate CDMA mobile phone speech. It makes use of the publicly available routines for the CDMA speech codec. Using this platform, speech data can be passed
through the codec under various modes of operation, while taking into account the under-
lying rules under which these might be initiated. This approach can encompass a large
number of possible scenarios in the network and makes it possible to relatively easily trans-
form any existing speech database into a CDMA-quality speech database. It also allows
studying the impact of different aspects of the network in isolation from other factors.

There are four key aspects of the CDMA mobile phone network which can directly impact
the speech signal. These are: (i) Dynamic Rate Coding (DRC), (ii) handling of Frame Loss
(FL), (iii) Background Noise (BN) at the transmitting end, and (iv) handling of Silence
Frames (SF). The latter aspect has not been investigated in this study because in FVC only
active speech is of interest. As far as possible, the impacts of the first three aspects have
been investigated in isolation. As will be explained, though, this is not entirely possible
because DRC is always occurring. But it is possible to constrain it to some extent.

With respect to the analysis technique used in this study to quantify the strength of speech
evidence, this thesis presents a new approach called Principal Component Analysis Kernel
Likelihood Ratio (PCAKLR). This is essentially an alternative to one of the commonly
used approaches, namely Multivariate Kernel Density (MVKD). It is shown that PCAKLR
exhibits similar FVC performance to MVKD for a small number of parameters. Most
importantly, though, it is computationally robust irrespective of the number of speech pa-
rameters used, an aspect of importance in terms of the speech parameter set used in this
research. PCAKLR also has a feature which allows it to handle within-segment as well
as between-segment correlations simultaneously. This provides an alternative way to fuse
results from multiple speech segments instead of using the standard logistic regression.

Among the various speech parameter sets commonly used in FVC, it is shown that Mel-
Frequency Cepstral Coefficients (MFCCs) are one of the best performing sets when dealing
with CDMA-quality speech. This is due to the fact that MFCCs are not a function of a
particular component of the speech production model that could be removed during the
CDMA coding process. Rather, they roughly estimate the energy in different frequency
bands of the speech signal.

As far as the impact of DRC on FVC analysis is concerned, surprisingly this is shown to
improve the accuracy and reliability of FVC analysis results when compared to uncoded
speech. It is argued that this improvement is linked to the quantisation process inherent in the speech codec which reduces within-speaker variation. With FL, this aspect is shown to negatively impact both same- and different-speaker comparisons of a FVC analysis when low speech coding quality is used. In the case of higher-quality speech coding, FL mainly negatively impacts different-speaker comparisons.

High levels of BN at the transmitting end of a CDMA mobile phone network, with Signal-to-Noise Ratios (SNRs) in the range of 9 to 15 dB, are shown to significantly worsen the accuracy of a FVC analysis. This is because the task of distinguishing BN from speech becomes a difficult task for the Noise Suppression (NS) process inherent in the CDMA speech codec, which begins to remove part of the original speech along with the BN present. However, the results of this study suggest that if a call is made from a highly congested area (e.g., a city centre), the negative impact of BN on FVC is likely to be less. This is due to the fact that when a large number of users access a cell site simultaneously, low speech coding quality is used to minimize the co-user interference in the CDMA network. This low-quality coding uses a different set of coding algorithms to the higher-coding qualities. As will be explained, specifically it repeats information from previous frames, which can mask some of the impact caused by BN.

In order to examine more realistic scenarios in the CDMA network, all the three aspects have been brought together and their impact on FVC assessed. It is shown that degradation in FVC accuracy results and this can be even more significant under mismatch conditions between the suspect, offender and background data. It is also shown that an improved accuracy can be obtained by passing the background data through the CDMA codec prior to FVC analysis. Though this goes a long way to mitigating the impact of the CDMA mobile phone network, it is still not as good as analysis under matched conditions using clean speech.
Acknowledgements

First and foremost, I must acknowledge and thank the Almighty Allah for blessing, protecting, and guiding me throughout this period. I could never have accomplished this without the faith I have in the Almighty.

I would like to express my deepest gratitude to my advisor Dr. Bernard John Guillemin for his excellent guidance, patience, confidence and helpful suggestions throughout the creation of this thesis.

I want to thank Dr. Catherine Watson for her helpful feedback and support.

I would like to thank my friend and research colleague, Balamurali Nair, who was always willing to help and share his ideas. The journal and conference publications of this research would not have been possible without his contribution.

I wish to thank the University of Auckland for funding my study and providing me with an excellent environment for doing research.

I would like to thank my proofreader, Catriona Carruthers, for her helpful comments and suggestions.

I express my deepest gratitude to my beloved mother, Sameera, and wish to tell her that her guidance and unconditional love have made me who I am today. She taught me to never give up and to follow my dreams. No matter how much I thank you, it will never be enough. Dad, may your blessed soul rest in peace.

Finally, I want to thank my two lovely sisters, Bayan and Aya, and my younger brother, Noor Aldin. They have always been a great support and encouraged me with their best wishes.
Contents

Abstract iii

Acknowledgements vi

Abbreviations 1

1 Introduction 3

1.1 Overview and objectives of this research 3

1.2 Thesis structure 10

1.3 List of publications 11

2 Background and Literature Review 13

2.1 Speech forensics and associated analysis methods 13

2.1.1 Auditory vs. acoustic analysis 13

2.1.2 Traditional acoustic parameters for Forensic Voice Comparison (FVC) 14
2.1.3 Automatic acoustic parameters for Forensic Voice Comparison (FVC) 14
2.1.3.1 Complex Cepstral Coefficients (CCCs) 15
2.1.3.2 Real Cepstral Coefficients (RCCs) 16
2.1.3.3 Linear Prediction Cepstral Coefficients (LPCCs) 16
2.1.3.4 Mel-Frequency Cepstral Coefficients (MFCCs) 16
2.1.3.5 Formant and Formant Trajectories (FT) 17

2.2 Expressing the outcome of an FVC analysis 18
2.2.1 Frequentist vs. Bayesian 18
2.2.2 The concept of likelihood ratio 19
2.2.3 Acceptance of the Bayesian approach 19
2.2.4 Approaches for estimating the likelihood ratio 20

2.3 Impact of mobile phone technologies on speech and FVC 22

2.4 Chapter summary 24

3 The Bayesian Likelihood Ratio Framework 26

3.1 Procedures for computing the likelihood ratio 26
3.1.1 Univariate Kernel Density (UKD) 27
3.1.2 Multivariate Kernel Density (MVKD) 29
3.1.3 Gaussian Mixture Model–Universal Background Model (GMM-UBM) 30
3.1.4 Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) 31
3.2 Ways of fusing the likelihood ratio results .................................. 35

3.2.1 Logistic regression ................................................................. 36

3.2.2 PCAKLR with No Fusion (PCAKLR_{NF}) ............................... 36

3.3 Tools for estimating the performance of an FVC analysis ................. 38

3.3.1 Numerical Methods ............................................................... 38

3.3.1.1 Validity (Log-likelihood-ratio Cost) ................................. 38

3.3.1.2 Reliability (Credible Interval) ........................................... 39

3.3.2 Graphical Methods .............................................................. 40

3.3.2.1 Tippett plots ................................................................. 40

3.3.2.2 APE plots ................................................................. 41

3.4 Chapter summary ................................................................. 43

4 Comparative Performance of MVKD vs. PCAKLR .......................... 45

4.1 Experimental methodology ...................................................... 45

4.2 Computational robustness of MVKD vs. PCAKLR in high-dimensional spaces ......................................................... 47

4.3 Performance of MVKD vs. PCAKLR for a small number of parameters 49

4.3.1 MVKD vs. PCAKLR for four parameters ................................. 49

4.3.2 MVKD vs. PCAKLR_{NF} for four parameters ............................ 51

4.3.3 PCAKLR vs. PCAKLR_{NF} for a large number of parameters ...... 54

4.4 Chapter summary ................................................................. 56
5  Impact of the CDMA Network on Speech Signals 57

5.1 Differences between communication technologies in respect of handling
the speech signal ................................................................. 57

5.1.1 Landline vs. mobile phone networks .................................... 57

5.1.2 The GSM vs. CDMA network ............................................ 58

5.1.3 Strategies for assessing the impact of mobile phone networks on
speech .............................................................................. 59

5.1.4 General description of the EVRC ........................................ 61

5.2 Ways in which the CDMA mobile phone network impacts on the speech
signal ............................................................................. 63

5.2.1 Dynamic Rate Coding (DRC) ............................................ 64

5.2.2 Frame Loss (FL) .......................................................... 68

5.2.3 Background Noise (BN) at the transmitting end ...................... 74

5.2.4 Silence Frames (SFs) ..................................................... 78

5.2.5 All aspects combined ................................................... 79

5.3 Chapter summary ............................................................ 80

6  The Comparative Performance of Automatic Acoustic Parameters for FVC 83

6.1 Experimental methodology .................................................. 83

6.2 Performance of uncoded speech ............................................ 87

6.3 Performance of EVRC-coded speech ..................................... 92

6.4 Chapter summary ............................................................ 100
7 Impact of Dynamic Rate Coding (DRC) on FVC Analysis

7.1 Experimental methodology ........................................ 101

7.2 FVC performance for low-, medium- and high-quality coded speech ........................................ 104

7.3 Chapter summary .................................................. 108

8 Impact of Frame Loss (FL) on FVC Analysis

8.1 Experimental methodology ........................................ 109

8.2 FVC performance for two codec modes under FL conditions ........................................ 113

8.3 Chapter summary .................................................. 117

9 Impact of Background Noise (BN) at the transmitting end on FVC Analysis

9.1 Experimental methodology ........................................ 119

9.2 FVC performance for two codec modes under various BN conditions ........................................ 121

9.3 Chapter summary .................................................. 130

10 Overall Impact of the CDMA Network on FVC Analysis

10.1 Experimental setup ................................................ 131

10.2 FVC Performance under matched/mismatched conditions ........................................ 134

10.2.1 Matched ..................................................... 134

10.2.2 Mismatched ................................................... 139

10.3 Comparison between impacts of the GSM and CDMA networks on FVC ............................. 143
10.4 Chapter summary ......................................................... 144

11 Conclusions and Directions for Future Work .................................. 146

11.1 Conclusions ................................................................ 146

11.2 Recommendations to forensic speech scientists ............................. 152

11.3 Directions for future work .................................................. 155

References ........................................................................... 157

Appendix A ............................................................................ 172

A Speech Parameters for Comparison ............................................ 172

A.1 Speech production and the source-filter model ............................... 172

A.2 Cepstral analysis .................................................................. 175

A.2.1 Complex Cepstral Coefficients (CCCs) .................................... 176

A.2.2 Real Cepstral Coefficients (RCCs) ........................................ 179

A.2.3 Linear Prediction Cepstral Coefficients (LPCCs) ........................ 179

A.2.4 Mel-Frequency Cepstral Coefficients (MFCCs) .......................... 180

A.3 Formants and Formant Trajectories (FTs) .................................... 181

Appendix B ............................................................................ 183

B The CDMA-Codec Platform ...................................................... 183

B.1 Platform usage for simulating mobile phone speech ........................ 183
B.1.1 Simulating the impact of DRC . . . . . . . . . . . . . . . 184

B.1.2 Simulating the impact of FL . . . . . . . . . . . . . . . 187

B.1.3 Simulating the impact of BN . . . . . . . . . . . . . . . . . 188
List of Tables

5.1 Equivalent mapping between speech and traffic coding bit rates 64

6.1 Performance of various speech parameters using uncoded speech 87

6.2 Performance of various speech parameters using EVRC-coded speech 93

7.1 Performance of different EVRC-coded speech qualities incorporating DRC 104

8.1 PESQ and MOS scores for a speech sample coded with OP0 at various percentages of FER 112

8.2 Impact of FL on FVC performance for the EVRC 113

9.1 Performance of FVC in the CDMA network using speech coded at OP0 and OP2 under various BN conditions 121

10.1 Transmission scenarios for the suspect, offender and background samples under matched and mismatched conditions 134

10.2 Performance of FVC in the CDMA network using medium-quality (MQ) speech coding, 10-15% FL under various BN levels with matched conditions 135
10.3 Performance of FVC in the CDMA network using medium-quality (MQ) speech coding, 10-15% FL under various BN levels with mismatched conditions.
List of Figures

2.1 The source-filter model ............................................. 16

3.1 Block diagram showing the arrangement of data in the PCAKLR approach . 34

3.2 Within-segment correlations for /a/ for (i) LPCCs, (ii) PCA transformed parameters; between-segment correlations for /a i e o u/ for (iii) LPCCs, and (iv) PCA transformed parameters. ............................................. 37

3.3 Tippett plot showing same- and different-speaker comparison results of an experiment. ............................................. 41

3.4 APE plots showing the losses in $C_{llr}$ from two different FVC experiments. . 42

4.1 $C_{llr}$ vs. number of input parameters for MVKD (red) and PCAKLR (green) for the speech segment /æ/ ............................................. 48

4.2 Tippett plot showing performance of PCAKLR vs. MVKD based on four speech parameters using conventional fusion. ......................... 50

4.3 APE-plot showing the $C_{llr}$ discrimination/calibration performance of PCAKLR vs. MVKD based on four speech parameters using conventional fusion. ............................................. 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>The performance of PCAKLR(_{NF}) vs. MVKD using 4 speech parameters.</td>
<td>53</td>
</tr>
<tr>
<td>4.5</td>
<td>The performance of PCAKLR(_{NF}) vs. MVKD using 4 speech parameters (zoomed version).</td>
<td>54</td>
</tr>
<tr>
<td>4.6</td>
<td>The (C_{llr}) discrimination/calibration performance of PCAKLR(_{NF}) vs. MVKD using 4 speech parameters.</td>
<td>55</td>
</tr>
<tr>
<td>4.7</td>
<td>(C_{llr}) vs. number of input parameters for PCAKLR(_{NF}) (red dots) and PCAKLR (blue dots).</td>
<td>56</td>
</tr>
<tr>
<td>5.1</td>
<td>Bit rate selection mechanism in the EVRC based on speech characteristics.</td>
<td>63</td>
</tr>
<tr>
<td>5.2</td>
<td>EVRC mapping of target ADR into different anchor operating points.</td>
<td>65</td>
</tr>
<tr>
<td>5.3</td>
<td>An example showing the switching mechanism between OP1, OP2 and vice versa in the case of coding voiced speech frames.</td>
<td>66</td>
</tr>
<tr>
<td>5.4</td>
<td>A set of time waveforms showing the impact of coding on the word “eight”. (a) Clean and uncoded speech, (b) speech coded at OP0, (c) speech coded at OP1, (d) speech coded at OP2.</td>
<td>67</td>
</tr>
<tr>
<td>5.5</td>
<td>A set of spectrograms showing the impact of coding on the word “eight”. (a) Clean and uncoded speech, (b) speech coded at OP0, (c) speech coded at OP1, (d) speech coded at OP2.</td>
<td>68</td>
</tr>
<tr>
<td>5.6</td>
<td>Illustration of the lost-frame mechanism implemented by the EVRC in the CDMA network. (a) A set of received speech frames, (b) Resulting set of speech frames used to reconstruct the speech signal.</td>
<td>72</td>
</tr>
<tr>
<td>5.7</td>
<td>A set of time waveforms produced for /aI/ coded with EVRC and corrupted at different frame locations. (Dashed lines show the frame boundaries).</td>
<td>73</td>
</tr>
</tbody>
</table>
5.8 Spectrograms of the time waveforms shown in Figure 5.7 for the EVRC.
   (Dashed lines show the frame boundaries). ........................................ 74

5.9 A set of time waveforms showing the impact of BN on the word “eight”. (a)
   Clean and uncoded speech, (b) uncoded speech with added street noise at
   SNR = 6 dB, (c) EVRC response to BN using OP0, (d) EVRC response to
   BN using OP2. (the red dashed lines show frame boundaries of the vowel
   /aI/ in eight). .................................................................................... 77

5.10 A set of spectrograms showing the impact of BN on the word “eight”. (a)
   Clean and uncoded speech, (b) uncoded speech with added Street noise at
   SNR = 6 dB, (c) EVRC response to BN using OP0, (d) EVRC response to
   BN using OP2. ................................................................................... 78

5.11 Impact of the CDMA network on speech. (a) processes implemented in
   sender mobile; (b) resulting processes in receiver mobile. In this figure it is
   assumed that Frames 6 and 7 are lost and that at the receiving end they are
   replaced by repeats of Frame 5. With respect to Frame 8 at the receiving
   end, the NS mechanism in built into the EVRC has removed part of the
   original speech signal. ................................................................. 81

6.1 Block diagram showing the experimental procedure used to investigate the
   performance of FVC using mobile phone speech ............................... 86

6.2 $C_{llr}$ v.s. $CI$ for uncoded speech for various speech parameters. ............ 88

6.3 Tippett plot showing FVC performance of CCCs using uncoded speech. .. 89

6.4 Tippett plot showing FVC performance of RCCs using uncoded speech. .. 89

6.5 Tippett plot showing FVC performance of MFCCs using uncoded speech. . 90
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Tippett plot showing FVC performance of LPCCs using uncoded speech.</td>
<td>90</td>
</tr>
<tr>
<td>6.7</td>
<td>Tippett plot showing FVC performance of FTs using uncoded speech.</td>
<td>91</td>
</tr>
<tr>
<td>6.8</td>
<td>APE plots showing the losses in $C_{llr}$ using various speech parameters extracted from uncoded speech.</td>
<td>92</td>
</tr>
<tr>
<td>6.9</td>
<td>$C_{llr}$ v.s. $CI$ for EVRC-coded speech at OP2 and OP0 for various speech parameters.</td>
<td>94</td>
</tr>
<tr>
<td>6.10</td>
<td>Tippett plot showing FVC performance of CCCs using high-quality speech coding (OP0).</td>
<td>95</td>
</tr>
<tr>
<td>6.11</td>
<td>Tippett plot showing FVC performance of RCCs using high-quality speech coding (OP0).</td>
<td>96</td>
</tr>
<tr>
<td>6.12</td>
<td>Tippett plot showing FVC performance of MFCCs using high-quality speech coding (OP0).</td>
<td>96</td>
</tr>
<tr>
<td>6.13</td>
<td>Tippett plot showing FVC performance of LPCCs using high-quality speech coding (OP0).</td>
<td>97</td>
</tr>
<tr>
<td>6.14</td>
<td>Tippett plot showing FVC performance of Formants using high-quality speech coding (OP0).</td>
<td>97</td>
</tr>
<tr>
<td>6.15</td>
<td>APE plots showing losses in $C_{llr}$ using various speech parameters extracted from speech coded at high quality (OP0).</td>
<td>98</td>
</tr>
<tr>
<td>6.16</td>
<td>APE plots showing losses in $C_{llr}$ using various speech parameters extracted from speech coded at low quality (OP2).</td>
<td>99</td>
</tr>
<tr>
<td>7.1</td>
<td>Block diagram illustrating how speech files have been processed corresponding to different capacity conditions.</td>
<td>102</td>
</tr>
</tbody>
</table>
9.3 $C_{llr}$ v.s. $CI$ for EVRC-coded speech at OP2 under various BN conditions.  . 124

9.5 Tippett plot showing FVC performance of speech uncorrupted with BN and coded at OP2 ................................................................. 125

9.6 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 9 dB coded at OP0. ........................................ 125

9.7 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 9 dB coded at OP2. ........................................ 126

9.8 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 15 dB coded at OP0. ........................................ 126

9.9 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 15 dB coded at OP2. ........................................ 127

9.10 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 21 dB coded at OP0. ........................................ 127

9.11 Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 21 dB and coded at OP2. ................................. 128

9.12 APE plots showing the losses in $C_{llr}$ using speech coded at OP0 under various BN conditions. ......................................................... 129

9.13 APE plots showing the losses in $C_{llr}$ using speech coded at OP2 under various BN conditions. ......................................................... 129

10.1 Block diagram showing the experimental procedure to assess the overall impact of the CDMA mobile phone network on FVC. ..................... 132
10.2 A bar graph showing the FVC performance ($C_{llr}$ and $CI$) for different scenarios in the CDMA mobile phone network under matched conditions.

10.3 Tippett plot showing FVC performance of uncoded speech using.

10.4 Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 9 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).

10.5 Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 15 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).

10.6 Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 21 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).

10.7 APE plots showing losses in $C_{llr}$ using speech coded under matched conditions.

10.8 A bar graph showing FVC performance ($C_{llr}$ and $CI$) for different scenarios in the CDMA mobile phone network under mismatched conditions.

10.9 Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 9 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario).
10.10 Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 15 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario). ................................................................. 141

10.11 Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 21 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario). ................................................................. 141

10.12 APE plots showing the losses in $C_{llr}$ using speech coded under mismatched conditions. ................................................................. 142

A.1 The source-filter model. ................................................................. 173

B.1 The EVRC codec platform layout and settings. ................................. 185

B.2 Block diagram illustrating how speech files have been processed corresponding to different capacity conditions. ........................................ 186
Abbreviations

AMR: Adaptive Multi Rate codec
AOP: Anchor Operating Point
APE: Applied Probability of Error
BN: Background Noise
CCC: Complex Cepstral Coefficient
CDMA: Code Division Multiple Access
CELP: Code Excited Linear Prediction
$C_{llr}$: Log-Likelihood-Ratio Cost
CI: Credible Interval
CN: Comfort Noise
DCT: Discrete Cosine Transform
DET: Detection Error Trade-off
DFT: Discrete Fourier Transform
EER: Equal Error Rate
EVRC: Enhanced Variable Rate Codec
FER: Frame Error Rate
FFT: Fast Fourier Transform
FL: Frame Loss
FVC: Forensic Voice Comparison
FT: Formant Trajectory
GMM-UBM: Gaussian Mixture Model–Universal Background Model
GSM: Global System for Mobile Communications
LLR: Log-Likelihood-Ratio
LR: Likelihood Ratio
LPC: Linear Prediction Coefficient
LPCC: Linear Prediction Cepstral Coefficient
MAP: Maximum a Posteriori
MFCC: Mel-Frequency Cepstral Coefficient
MOS: Mean Opinion Score
MVKD: Multivariate Kernel Density
NELP: Noise Excited Linear Prediction
NS: Noise Suppression
PAV: Pool Adjacent Violators
PCA: Principal Component Analysis
PCAKLR: Principal Component Analysis Kernel Likelihood Ratio
PCAKLR_{NF}: Principal Component Analysis Kernel Likelihood Ratio-No Fusion
PDF: Probability Density Function
PESQ: Perceptual Evaluation of Speech Quality
PPP: Pitch Period Prototype
RCC: Real Cepstral Coefficient
SF: Silence Frames
UA: Univariate Analysis
UKD: Univariate Kernel Density
Chapter 1

Introduction

1.1 Overview and objectives of this research

The use of mobile phone technology has dramatically increased over the last decade. Not surprisingly, recordings of mobile phone conversations are increasingly being used as evidence in courts of law. Analysis of such recordings using a range of Forensics Voice Comparison (FVC) techniques can assist the court in establishing the guilt or innocence of a suspect. However, forensic speech scientists, when undertaking such analysis, may assume that analysis techniques used for speech transmitted over one mobile phone network are applicable to speech transmitted over another [1, 2, 3, 4]. The most widely used mobile phone technologies in use today are the Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). These two network topologies are fundamentally different in their design and internal operation. Given these major differences, the ways in which they impact on the speech signal, usually negatively from the perspective of forensic analysis, are also very different.

The primary goal of this thesis is to investigate the direct impact of key features of the CDMA mobile phone network on the speech signal and thus the outcome of an FVC analysis. An understanding of these features is essential in order to consider in what ways, and to what extent, the speech signal might be negatively impacted by such a network. These features are: (i) Dynamic Rate Coding (DRC), (ii) strategies for handling Frame Loss (FL),
1.1 Overview and objectives of this research

(iii) strategies for overcoming the effects of Background Noise (BN) at the transmitting end, and (iv) handling Silence Frames (SF). In this study, only the first three aspects and their respective impacts on FVC have been investigated. This is because in a real forensic casework only active speech is used in the comparison process. Analysis of SF is still important, though, to forensic scientists. SF present in a mobile phone call carries information about BN at the sending end. When no significant changes of BN energy levels are detected at the sending end, the last received SF is repeated [5, 4]. This fact, for example, can be utilized to determine if the original recordings have been tampered with prior to the analysis. Before discussing the specifics of these features, as will be explained later in Chapter 5, it is important at the outset to appreciate how mobile phone networks are different with respect to these features.

One key differences between the GSM and CDMA networks is with respect to the mechanism driving the process of dynamically changing the source coding bit rate. With the GSM network, changing channel conditions, referred to as channel quality, is the driver; with the CDMA network, it is changing user demand, referred to as channel capacity. The source coding bit rate directly impacts on the resulting quality of the coded speech signal. Mobile phone networks incorporate highly sophisticated speech coding blocks, called codecs, which code the original speech in order to achieve a reasonable level of data compression (i.e., low bit rate). The most widely used speech codecs in the GSM and CDMA networks are the Adaptive Multi Rate (AMR) codec and the Enhanced Variable Rate Codec (EVRC), respectively. These codecs have many modes of operation, which in turn govern, among other things, the resulting bit rate. The network initiates changes between these modes and this is referred to here as dynamic rate coding.

Another important aspect that cannot be overlooked in these networks is with respect to the medium of transmission (i.e., the wireless channel). This can often be quite poor, increasing the likelihood of frames being lost or corrupted during transmission. In order to maintain good quality speech under such conditions, innovative techniques have been implemented in the speech codecs of both the GSM and CDMA networks to mitigate the impact of this on speech quality [5, 6]. Specific strategies for achieving this in the GSM and CDMA networks differ, but there are some common features. Speech is coded into 20
ms frames and, when a speech frame is received, certain bits are checked by the speech codec to determine whether errors in transmission have occurred. If they have, there is an attempt to correct them using, for example, convolutional coding [7]. If this process fails, or indeed if a frame is lost completely, a new frame will be created to replace it. The data in this frame could be either a repeat of a prior undamaged speech frame, or an extrapolation from undamaged frames already received.

Another important difference between the GSM and the CDMA networks is with respect to the process of handling BN at the transmitting end. The EVRC incorporates a special feature called Noise Suppression (NS), which is not present in the GSM network. NS uses a sophisticated mechanism to reduce BN prior to analyzing and encoding the speech signal. This process helps to improve the distinction between speech and BN for the purpose of detecting silence frames. It also helps to improve the classification of speech frames as voiced, unvoiced or transient; parameters which are necessary for the dynamic rate coding processes implemented by the EVRC.

The only component in the CDMA network which directly handles the speech signal, and is thus responsible for any changes that might occur to it, is the EVRC speech codec. The EVRC has many modes of operation, as will be discussed later in this thesis. Changing which mode is in operation at any particular time is a process controlled and initiated by the network as a whole in response to changing capacity conditions, or by the user’s mobile phone in response to changing speech characteristics. Therefore, if one understands all of the EVRC’s possible modes of operation and, the constraints with respect to how these modes can be changed, this will encompass the majority of impacts on the speech signal under most transmission scenarios. This strategy is much better than conducting a large number of experiments involving transmission of speech across an actual mobile phone network. This latter strategy is likely to reveal only a small subset of possible impacts on the speech. Further, there would be no way of knowing which possible transmission scenarios had been represented because such information is not available in the mobile phone recording (i.e., the received speech signal) [1]. With the first approach in mind, a new software platform, namely the EVRC-codec platform, has been developed in order to
simulate typical scenarios in the CDMA mobile phone networks and their resulting impacts on the speech signal.

In terms of number of users and geographical coverage of the CDMA network, recent surveys have reported that the total number of users utilizing the CDMA technology worldwide is approximately 500 million, compared to 3 billion using the GSM network. The GSM network is more popular worldwide, with operators in 212 countries [8]. Nonetheless, the CDMA network is still very popular in North America, China and India with a presence in 118 countries worldwide [9].

The evaluation of forensic speech evidence in this study is undertaken using the Likelihood Ratio (LR) framework. This framework usually consists of two major concepts: similarity and typicality, the former quantifying the amount of difference between two speech samples, the latter their typicality with respect to a relevant background population. Within the LR framework different methods have been established to evaluate speech evidence such as Univariate Analysis (UA), Univariate Kernel Density (UKD), Multivariate Kernel Density (MVKD) and Gaussian Mixture Model-Universal Background Model (GMM-UBM) analyses [10, 11, 12].

With UA and UKD, individual LRs are computed for each parameter and then, under the assumption of independence between input parameters, these are combined naïvely via simple multiplication to produce an overall LR value (i.e., the naïve Bayes concept [13]). Given that some measure of correlation (i.e., dependency) between parameters is to be expected [11], the result is often an over- or under-estimation of the strength-of-evidence. UKD, however, is preferred over the UA because it uses a kernel density distribution instead of a normal distribution to more accurately model the background population. It also has the advantage of being simple to implement and was found to be computationally robust.

The MVKD estimation of the LR was developed by Aitken and Lucy in the context of the analysis of elemental glass fragments [10]. This method has grown in popularity in the FVC arena because of its ability to handle correlations between multiple input parameters from the same speech segment. MVKD was originally proposed for use with a small number of parameters. However, some researchers have started using it for larger numbers of parameters, in the region of 14. Promising results have been presented using this approach.
1.1 Overview and objectives of this research

[14, 15], and it was initially intended for use in this research. However, in the early stages of this study, this approach was found sometimes to produce invalid results when used with a larger number of parameters than it was originally designed for. The cause of this fragility, as will be discussed in Chapter 3, is linked to the difficulty of smoothing the kernel density function in high dimensional spaces, as well as the inversion of ill-conditioned matrices, which become even more problematic as matrix size increases.

The GMM-UBM approach is commonly used in automatic speaker recognition [16, 17], but has also been successfully applied to FVC [18]. However, it typically requires a very large amount of data for determining the background population model. Given that in practical forensic scenarios the amount of data available for analysis is often limited, this approach has not been considered in this study.

In this thesis, a new approach is presented for computing LRs based on principal component analysis (PCA). This has been found to produce good results, even when the number of input parameters is large. With this approach, termed Principal Component Analysis Kernel Likelihood Ratio (PCAKLR), multiple input parameters are first transformed into sets of uncorrelated parameters [19, 20]. The resulting principal components are guaranteed to be orthogonal provided that the input data is jointly normally distributed [20, 21]. An LR is then computed individually for each of these sets using UKD and their product computed to produce an overall LR value according to the naïve Baye’s approach. PCAKLR shares many features with MVKD in that: (i) the suspect data is modelled using a normal distribution, (ii) background data is modeled using a set of kernels and (iii) correlations are taken into account in the comparison process. The key difference between these approaches is that PCAKLR is mathematically robust irrespective of the number of input parameters.

Another interesting feature of PCAKLR is the manner in which LR results can be combined for multiple speech segments (e.g., vowels). In the case of MVKD, for example, this is traditionally done using logistic regression fusion, which accounts for the correlation between LRs from multiple segments. Though the same approach can be used for PCAKLR, an alternative strategy is to bring together the parameter sets for individual vowels into one large data matrix and leave it to the PCA stage of PCAKLR to remove all correlations. This
1.1 Overview and objectives of this research

is referred to as PCAKLR\textsubscript{NF} (NF: No Fusion). The comparative performance of MVKD, PCAKLR, and PCAKLR\textsubscript{NF} is discussed in Chapter 4.

The comparison process in this study has used formant trajectories and different kinds of cepstral coefficients, namely Complex Cepstral Coefficients (CCC), Real Cepstral Coefficients (RCC), Linear Prediction Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC). The best performing parameter set among these has been determined for the CDMA network on the basis of the accuracy and precision of LR results. The chosen parameter set, as shown in Section 6.3, was then used in the subsequent experiments when assessing the impact of different aspects of the CDMA network on FVC.

It is known that cepstral coefficients are generally sensitive to transmission artefacts and various techniques have been proposed to compensate for this [22, 23]. Though this is an issue when working with landline phone recordings, the manner in which transmission artefacts impact on the speech signal in mobile phone networks is entirely different. For example, speech data is transmitted in packets. If a packet gets lost or corrupted during transmission, this will be detected by a sophisticated error detection mechanism and there will be an attempt to correct it. If the correction is not possible, a new packet will then be inserted using information from previous good speech frames [18]. Therefore, partially corrupted speech data as a result of, for example, channel noise, never arrives at the receiving end. Thus the compensation and normalization techniques normally used for landline networks are not appropriate when working with mobile phone speech. Distinction must also be made between channel noise and the kind of noise present at the transmitting end (i.e., BN). The latter has a direct impact on the speech signal and it is one of the key aspects being investigated in this study.

A number of speech databases have been used in these experiments, namely the TIMIT, NRIPS and XM2VTS databases. The Timit database was only used in the early stages of this research for developing PCAKLR and fine tuning its performance [24]. The sole purpose of these experiments was to examine the behaviour of PCAKLR under conditions where MVKD cannot be guaranteed to produce accurate results (i.e., when the number of input parameters is well in excess of three or four). In these experiments, vowel phonemes were extracted from the speech of 11 male speakers of a North American dialect, and the
number of input parameters was sequentially increased from 1 to 14. LR values were then computed using the MVKD and PCAKLR approaches and the resulting FVC accuracy was compared for each case. Though it is acknowledged that the Timit speech data is far from being forensically realistic data (as it contains clean read speech and does not include non-contemporaneous recordings) these experiments did serve to demonstrate the robustness of PCAKLR and the fragility of MVKD when the number of input parameters is larger than MVKD was originally designed for.

In order to assess the performance of MVKD, PCAKLR, as well as PCAKLR_{NF}, using data that is more realistic, the NRIPS database was used. This was collected by the Japanese National Research Institute of Police Science for forensic speaker recognition tests. It includes recordings of excerpts of the five standard Japanese vowels (/a i e o u/) spoken by 297 male Japanese speakers and transmitted over a landline telephone network. The tokens were extracted from read speech and includes non-contemporaneous recordings separated in time by two to three months [25] (see Chapter 4 for further details).

Finally, the XM2VTS database was used in the majority of these experiments to investigate the impact of the CDMA mobile phone network on the speech signal and thus the performance of FVC analysis. 130 male speakers of this database were considered in these experiments [26]. These speakers had been recorded on four different occasions at intervals of one month and during each session each speaker had read three sentences twice. Speech files of the 130 speakers were then coded in this study using the EVRC-codec platform under various channel and capacity conditions to reflect a wide range of scenarios in the CDMA network.

It needs to be noted here that having a mismatch between the recordings of offender, suspect and relevant background population represent a common scenario in forensic casework [27, 28, 29]. Forensic speech scientists often deal with various types of mismatch conditions, including, but not limited to: different sample sizes (i.e., the number of tokens available for each speaker), dialects mismatch (e.g. background data contains a mixture of dialects that are different to the suspect and offender dialect) and different transmission channels. For example, it is typical for the offender speech data to have been acquired from mobile phone recordings, and therefore to be of relatively poor quality, while the suspect speech
data has come from recordings of police interviews and to be of relatively good quality. Though this aspect of mismatch conditions is not the focus of this study, but one example of mismatched conditions have been considered at in Chapter 10. In this example, the suspect and offender speech data were assumed to be of CDMA mobile phone quality and background data was clean uncoded speech. The rationale being that the background data is typically collected in a recording studio, resulting in a high-quality speech. (Note: the EVRC-codec platform developed as part of this research makes it possible to transform, studio-quality recordings into mobile-phone quality recordings [2]). In this way, mismatch between offender, suspect and background speech samples can be reduced.

1.2 Thesis structure

The remainder of this thesis is structured as follows. Chapter 2 discusses previous work and the current state-of-the-art in speech forensics. Chapter 3 overviews the Bayesian likelihood ratio framework and the different analysis techniques currently used for computing an LR. It also demonstrates how the PCAKLR technique, which was developed as part of this study, should be implemented when conducting an FVC analysis. This chapter also describes the different tools and metrics used in this study to investigate the manner and extent to which the outcome of an FVC analysis can be impacted by the CDMA mobile phone network. These tools are Log-likelihood-ratio Cost ($C_{llr}$), Tippett plots, Credible Interval ($CI$) and Applied Probability Error Plots (APE). Chapter 4 examines the cause of fragility of the MVKD algorithm and then compares its performance to PCAKLR and PCAKLR$_{NF}$. Chapter 5 provides background information necessary to understanding the different aspects of the CDMA mobile phone network that can directly impact the speech signal. It comprises two major sections. Section 5.1 overviews the differences between different communication technologies, specifically the landline network and GSM and CDMA mobile phone networks. This also includes a general description of the speech codec used in the CDMA network. Section 5.2 discusses key features of the CDMA mobile phone network which can negatively impact the speech signal, these being: (i) Dynamic Rate Coding (DRC), (ii) mechanisms for handling Frame Loss (FL), (iii) handling of Background Noise.
1.3 List of publications


- Nair, B. T., Alzqhoul, E. A., and Guillemin, B. J. "Comparison between Mel-Frequency and Complex Cepstral Coefficients for Forensic Voice Comparison using mobile phone speech data".

---

1In this example mobile-phone quality is assumed for the suspect and offender speech data, and studio-quality for background data.


Chapter 2

Background and Literature Review

2.1 Speech forensics and associated analysis methods

There is an increasing demand in the legal arena for the opinion of speech scientists as to whether two or more speech recordings are from the same or different speakers. This process is usually termed Forensic Voice Comparison (FVC) and it has been used for quite some time. It was first implemented by a German institution, Bundeskriminalamt, in 1980 [30, 13]. There is, however, no consensus of opinion among speech scientist as to how FVC analysis should be implemented in practice. The most popular analysis methods, though, are: (i) Auditory Phonetic and Acoustic Phonetic Analysis, which uses a combination of acoustic and auditory parameters, and (ii) Automatic Speaker Recognition with Human Analysis (HASR), which uses an automatic system to find the similarities between speech samples and compares them with respect to their auditory and acoustic properties [31, 32, 33, 34].

2.1.1 Auditory vs. acoustic analysis

With auditory analysis, the expert compares speech and voice samples on the basis of how they sound. With respect to acoustic analysis, the expert compares speech samples in terms
of their acoustic properties, typically extracted by computer software. Both types of parameters are equally important, but logically, the auditory analysis precedes acoustic analysis, as it is necessary to listen to the samples in question to first decide whether one should proceed further with the analysis or not. When listening to speech samples, for example, one might note that a certain vowel, in similar phonological context, is contained in many words and the acoustic parameters of this could then be used in the comparison process [13, 34].

2.1.2 Traditional acoustic parameters for Forensic Voice Comparison (FVC)

Traditional acoustic parameters are those that signal the differences between different languages or even within a language itself [35]. For example, linguists understand the difference between the vowel /ea/ in “head” and the vowel /a/ in “had”. It is also known that pitch period relates to fundamental frequency (F0) and, due to the different physical structures of speakers’ vocal folds, the same sound can be produced at different values of F0 for different speakers. A considerable amount of research has focused on these types of parameters and they are well understood by the linguistic and phonetician communities.

2.1.3 Automatic acoustic parameters for Forensic Voice Comparison (FVC)

Automatic acoustic parameters, on the other hand, are difficult to conceptualise and relate to the articulatory properties of speech sounds. For example, while pitch can be related to the traditional acoustic parameter (F0), the fifth or sixth cepstral coefficients are difficult to relate in a straightforward way to something phonetic. Nevertheless, automatic acoustic parameters are widely used in FVC analysis and have been used in this investigation for a number of reasons: (i) they can be automatically extracted, (ii) large numbers of parameters can be determined, (iii) the process is less labour intensive, and (iv) they have been shown to
outperform traditional parameters in many applications [36, 37, 13]. Examples of these parameters are: (i) the Complex Cepstral Coefficients (CCCs), (ii) Real Cepstral Coefficient (RCCs), (iii) Mel-Frequency Cepstral Coefficients (MFCCs), and (iv) Linear Prediction Cepstral Coefficients (LPCCs). Though formants and formant trajectories are classified as traditional acoustic parameters, and normally extracted using human-supervised tracking methods, they can also be automatically extracted. The human-supervised methods are considered more reliable and accurate than the automatic methods when used with high-quality speech recordings, but are very labour intensive. However, both methods have been shown to give comparable performance when used with telephone-quality speech [27]. Therefore, only automatic methods have been considered in this research. The next sections provide an overview of the different types of automatic acoustic parameters used in this study (note: further details about the various types of cepstral coefficients used as well as the speech production model can be found in Appendix A).

2.1.3.1 Complex Cepstral Coefficients (CCCs)

CCCs are the output from the standard form of cepstral analysis, which is often referred to as homomorphic filtering [38]. This is used to separate out the various aspects of the speech production model, which is comprised of: (i) the excitation source $P(Z)$ (periodic in the case of voiced sounds), (ii) glottal shaping filter $G(Z)$, (iii) vocal tract filter $V(Z)$ and (iv) lip radiation filter $R(Z)$. Cepstral analysis is defined as the inverse Fourier transform of the natural logarithm of the Fourier Transform (FT) of the speech signal [39] (see Figure 2.1). When this process is applied to the magnitude and phase components of speech, the resulting set is called the complex cepstrum. The glottal shaping filter is an anti-causal filter with zeroes typically outside the unit circle and poles located at the origin of the $Z$-plane. The vocal tract filter is typically an all-pole filter with all of its poles inside the unit circle. The lip radiation filter typically has only a zero inside the unit circle and a pole at the origin [39]. Therefore, the resulting complex cepstrum is comprised of both causal and anti-causal components, the causal components arising from the vocal tract and lip radiation filters, and the anti-causal components arising from the zeros of the glottal shaping filter.

\footnote{A is the gain of the glottal shaping filter}
The filter components of speech are dominant in the lower part of the complex cepstrum, whereas the excitation source dominates the upper part.

### 2.1.3.2 Real Cepstral Coefficients (RCCs)

The process of extracting RCCs differs slightly to that used for CCCs. Specifically, the logarithm operator is applied to the magnitude of the FFT component only, and phase information is ignored. The term cepstrum is typically used for RCCs and they represent the even component of the CCCs [40, 39]. Therefore, RCCs inherently carry less information.

### 2.1.3.3 Linear Prediction Cepstral Coefficients (LPCCs)

The linear prediction cepstral coefficients are derived directly from LPC coefficients. LPC analysis assumes a simplified source-filter model which lumps together the vocal tract, lip radiation and glottal shaping filters and assumes that the resultant filter is an all-pole with poles inside the unit circle. Consequently LPCCs contain only the causal components of speech [41]. The simplified LPC model, together with the fact that the zeroes of the speech production model are ignored, suggest that LPCCs do not carry as much speaker specific information as either CCCs or RCCs. They could be, therefore, expected to perform less well in FVC.

### 2.1.3.4 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are a perception-based parameter set. The speech signal is first transformed into the frequency domain using the Discrete Fourier Transform (DFT). The next step is to
2.1 Speech forensics and associated analysis methods

estimate the amount of energy existing in various regions of the frequency domain. This is estimated over a set of overlapped Mel-filter banks by computing the power spectrum of the speech signal and then summing up the energies in each filter bank region. Once energies of these filter banks are computed, the logarithm operator is applied. Finally, a Discrete Cosine Transform (DCT) is applied to the logarithm of the energies, which results in a set of MFCC coefficients [38]. Although the overall process of MFCC extraction aligns with the procedure for cepstral analysis, it is arguable whether the MFCCs should be classified as cepstral coefficients. This is because MFCCs do not separate out any components of the speech production model, as is the case for other types of cepstral coefficients. It is a common practice to use deltas and delta-deltas along with MFCCs to capture the dynamic aspect of the speech signal [1]. These are simply the first and second order derivatives of the MFCCs over a range of short-term speech frames.

2.1.3.5 Formant and Formant Trajectories (FT)

Speech formants are classified as traditional acoustic parameters, as they can be interpreted by linguists and linked to the physical shape of the vocal tract. However, these have been automatically extracted in this thesis and therefore have been included here. Formants are sometimes referred to as the acoustic resonances of the vocal tract, and are identified by locating dominant peaks of the speech spectral envelope [42]. A change in formant value from one speech frame (or segment) to another constitutes a formant trajectory. Formant trajectories capture the dynamic aspects of speech and are widely used in speech applications, such as speaker and speech recognition. with respect to the experiments conducted in this study, only the first three formants were considered. The rationale being that speech codec in the CDMA network band limits the signal between 200 Hz and 3.4 kHz. Therefore, any information related to the fourth and fifth formants, for example, will either be lost or highly distorted. Discrete cosine transforms were then fitted to the formant trajectories and only the first four DCT coefficients were used. In the experiments of this study, it has been found that using DCT coefficients higher than four would only add a little improvement to the trajectory representation. Further, DCTs in the range of four or five have been typically used in previous studies such as the work in [27].
2.2 Expressing the outcome of an FVC analysis

2.2.1 Frequentist vs. Bayesian

There are two major schools of thought as to how speech evidence should be legally expressed and presented to a court of law, these are: the Frequentist and Bayesian schools. Frequentists adopt a variety of frameworks such as the Classical Probability Scale (CPS) and UK Position Statement\(^2\) [43, 44, 34]. These frameworks generally employ the combined auditory and acoustic phonetic approach. They also involve assessment for the probability of differences in the measurement of acoustic parameters between two speech samples (e.g. suspect and offender). Subsequently, the suspect sample would be considered a “match” if it falls within a certain defined distance (or threshold) of the offender sample.

There are a number of concerns associated with the Frequentists approach. First, it involves a somewhat binary decision, where a suspect would be eliminated if voice samples differed by slightly more than the selected threshold, but would be considered a “match” if the difference was slightly below this threshold [13, 45]. Second, it is difficult to select an appropriate threshold when the within-speaker variation is higher than the between-speaker variation, which is a situation typically encountered in speech analysis. The improper selection of a threshold can result in a large number of different-speaker comparisons being falsely identified as same speaker and vice versa.

The Bayesian school adopts the Likelihood Ratio (LR) framework for presenting speech evidence [46, 47, 48, 49]. With this framework, the strength of speech evidence is the quantified degree of similarity or typicality between the speech samples under investigation on the basis of two hypotheses. The first is that these samples come from the same speaker and this is normally referred to as the prosecution hypothesis \(H_P\). The second, the defence hypothesis \(H_d\), often states that samples were spoken by different speakers, who happened to sound the same [50, 51]. Though, the LR is a number, some experts express it verbally, as its numerical form may not be readily interpretable by a court [52]. More recently, some

\(^2\)It needs to be noted here that the UK position approach is not strictly a frequentist approach. It was developed to overcome some issues with the CPS and can be considered as transition between the frequentist and the likelihood ratio approach.
practitioners in the UK have stopped using the UK position statement and now using a
verbal likelihood ratio and support statement.

2.2.2 The concept of likelihood ratio

The LR measures the probability of evidence given two competing hypotheses – the pros-
ecution (speech samples have the same origin) and defence (speech samples have different
origins). Mathematically the LR is defined as: \( \text{LR} = \frac{p(E|H_p)}{p(E|H_d)} \), where \( p(x|y) \) is the con-
ditional probability of \( x \) given \( y \), \( E \) is the evidence, and \( H_p \) and \( H_d \) are the prosecution
and defence hypotheses, respectively. LR values significantly greater than 1 support the
prosecution hypothesis, values significantly less than 1 support the defence hypothesis, and
values close to 1 provide little evidence either way. LRs are often converted into a loga-
rithmic scale (i.e., Log-Likelihood-Ratio \( \text{LLR} = \log_{10}(\text{LR}) \)). This is thought to be an easier
metric to use, where large positive LLRs support the prosecution hypothesis, large negative
values support the defence hypothesis, and values close to 0 support neither prosecution
nor defence [53, 54].

2.2.3 Acceptance of the Bayesian approach

The evaluation of forensic speech evidence using the LR framework has gained a greater
acceptance among forensic speech scientists over the last decade [53, 34]. The most impor-
tant advantage of the Bayesian approach is that it a logically correct theoretical framework.
It has also been shown to work when applied to forensically realistic material such as DNA,
speech and glass fragments [1, 55, 56]. The Bayesian approach predicts that the same- and
different-subject data are resolved with LRs greater and less than 1, respectively. Another
important feature of this approach is that it makes combining separate pieces of evidence
straightforward provided they are mutually independent [45, 13]. Therefore, a set of LRs
computed from different independent sources or pieces of evidence can be combined by
taking their product.

As far as FVC analysis is concerned, recent studies have demonstrated the effectiveness
of this approach. One study in 2001 [56], which used Japanese speech data, has
shown that around 90% of the same- and different-speaker comparisons were resolved with LRs greater and less than 1, respectively. Two other studies that used American English [57] and Spanish [58] also obtained a classification rate of about 86% for both the same- and different-speaker comparisons. However, the classification-error rate and correct-classification rate [59, 60] used in these studies for assessing the performance of an FVC analysis have become less popular, with the development of new metrics that are more appropriate when working with the LR framework. One issue with the more traditional metrics is that they involve a binary decision, where errors are counted when, for instance, the FVC system declares two samples to come from a different origin, when in fact they have the same origin (i.e. a false negative). Further, they are based on posterior probabilities, which is the trier-of-fact’s belief after presenting the evidence. This is not within the scope of forensic scientists, as they should only focus on estimating the strength of the evidence (or the LR calculation) based on the samples available to them. The newer metrics used currently are: the Log Likelihood-Ratio Cost ($C_{llr}$) for estimating the accuracy (or validity) [61, 62], and the Credible Interval ($CI$) for estimating the precision (or reliability) [63] of an FVC analysis. The increasing popularity of these metrics is evidenced by the fact that they are used in a large number of recent publications [64, 65, 66, 67, 68, 15]. Further details about these metrics are provided in Chapter 3.

### 2.2.4 Approaches for estimating the likelihood ratio

Different procedures for calculating the LR have appeared in the literature. One of the earliest procedures was developed by Lindley [69] for univariate data. More recently, several procedures were presented by Aitken and Lucy [10], which operate directly on the multivariate data and account for the within- and between-speaker variations. Some of these procedures assume a normal distribution for the between-speaker data (or background data), while others uses a sum of equally-weighted kernels. Each one of those kernels is a Gaussian distribution centred on the mean value of each subject of the background data. This latter approach is termed Multivariate Kernel Density (MVKD) and has been found to outperform the univariate procedures [10]. MVKD has gained wide acceptance among forensic scientists and it is commonly used with traditional acoustic parameters, such as
formants and formant trajectories [70, 15, 66]. However, it has also been used with a large number of automatic acoustic parameters, showing promising results [15, 71, 14, 72]. However, as will be illustrated in Chapters 3 and 4, one cannot guarantee that MVKD will always provide accurate results when used with a large number parameters (i.e., more than the number it was originally designed for).

Another approach, which is common in automatic speaker recognition and has been recently applied to FVC and shown good results, is the Gaussian Mixture Model–Universal Background Model (GMM-UBM). The main difference between the GMM-UBM and MVKD is with respect to modelling the background data. The Universal Background Model (UBM) is trained on all background data pooled across speakers and normally requires a large amount of data to achieve good performance.

One recent study [15] compared the performance of GMM-UBM with the MVKD analysis using DCT coefficients of the second formant trajectory extracted from a number of phonemes. This study has shown that both MVKD and GMM-UBM had similar performance in terms of the accuracy of LR results when computed for individual phonemes. However, when LRs from all phonemes were fused, using a standard procedure called logistic regression, the GMM–UBM system outperformed MVKD analysis. In terms of the precision of FVC analysis, it showed that MVKD tends to have better precision for the low LR values, but that this precision drops off as the LR values increase. However, the overall mean CI value was shown to be better for the GMM-UBM model.

The performance of MVKD has also been compared to the new approach developed in this study for calculating LRs, namely Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [68] (details of PCAKLR can be found in Chapter 3). With this approach, the multivariate data of the suspect, offender and background population are first decorrelated and transformed into a new vector space. As with the MVKD approach, the background data in PCAKLR is modelled using a set of Gaussian distributions of equal weights centered on the mean value of each subject. An LR value is calculated for each transformed parameter using Univariate Kernel Density (UKD) analysis and then the product of these individual LRs is taken to produce an overall LR value. The comparison between these two procedures is discussed in Chapter 4.
2.3 Impact of mobile phone technologies on speech and FVC

Forensic speech scientists may erroneously assume that the analysis techniques used for landline-quality speech are applicable to speech transmitted over mobile phone networks. As will be explained in Section 5.1, these networks are fundamentally different in their ways of coding, filtering, digitizing and transmitting the speech signal. Further, within the mobile phone arena there are two different topologies, these being: the Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). These networks differ fundamentally in their methods of handling and processing the speech signal, and thus their impact on the outcome of an FVC analysis can be expected to differ.

One main difference between the landline and mobile phone networks is with respect to the medium of transmission, which is wireless in the latter. The characteristics of this wireless channel can change significantly over short periods of time and are subject to many external factors such as user mobility and weather changes, etc. This can result in fading, interference, channel noise and channel phase distortion which are likely to impact the transmission of speech. However, the speech signal itself is not directly impacted by such factors, but rather in a highly indirect manner. When channel conditions are bad, irrespective of the cause, one of the following could happen to the transmitted speech frames: (i) they are lost, (ii) they are irrecoverably corrupted, or (iii) they are partially corrupted and are then subsequently repaired using sophisticated error detection and correction routines [4]. In all cases the outcome is the same, namely ‘clean’ speech, though in the first two of these the ‘clean’ speech will be generated in some manner from uncorrupted speech data from the past. The network also responds to bad channel conditions and increasing frequency interference by sending a set of instructions to the speech codec. Mobile phone networks incorporate highly sophisticated speech coding blocks, called codecs, which code the original speech in order to achieve a reasonable level of data compression (i.e., low bit rate). The network instructions to change the codec modes of operation are used to minimize the impact of adverse conditions and thus increase the robustness of transmission (see Chapter 5 for details). Though the result of changing the codec mode of operation (or
source coding bit rate) is also clean speech, this aspect can directly impact on the resulting quality of coded speech. It needs to be noted here that cepstral coefficients, for instance, are known to be generally sensitive to transmission artefacts and various techniques have been proposed to compensate for this [22, 58]. Though this is an issue when working with landline phone recordings, the same is not true for mobile phone recordings.

There have been a few studies examining the impact of the GSM network on the speech signal, but the author is not aware of similar studies focusing on the CDMA network. In 1997 a collaborative study investigated the impact of different coding schemes, including an early version of the speech codec used in the GSM network, on speaker identification [73]. With respect to the GSM codec, this study showed an impact on formant frequency trajectories for different speech segments, especially those involving rapid formant transitions. Another study in 2007 examined the overall impact of the GSM network on formant frequencies [74]. This showed considerable differences in formant frequency measurements between the coded and uncoded speech. Also, measurements of the first formant were observed to be higher, though it is possible that this in part was caused by the filtering of the speech signal below 300 Hz, which precedes the GSM coding. Then in 2008 a study looked at the impact of the GSM network on formant frequencies as well as pitch [4]. It concluded that though the impact on pitch is minimal, the impact on measured formant frequencies can be quite significant. It also showed that at times the codec can change unvoiced speech into voiced, and vice versa. A study in 2010 [75] also examined the impact of landline recordings, which originated from a cellular phone on the GSM network, on the measurement of formants. It revealed that a major contributor to changes in the measurement of the first formant is with respect to the band-pass filtering process introduced by the landline network. However, the same study noted that the impact of the GSM codec itself was rather small.

Another more recent work in 2011 [76] investigated the impact of a number of speech codecs, including the GSM codec, on the performance of a speaker recognition system. The speech data was passed through the speech codec using various fixed bit rates. It concluded that degradation to the accuracy of comparison results was negligible for most of the codecs operating at high bit rates (around 15 kbps). In respect of the GSM codec, the
results showed a drop in the performance of comparison results at low bit rates, which can be as high as 20%. However, this performance improves as the bit rate increases. This same study also examined the impact of packet loss on speaker recognition using a number of Voice over IP (VoIP) codecs such as SILK. Again, only a slight drop in the performance was observed with packet losses up to 20%. A very recent study in 2013 [27] also examined the impact of telephone transmission, including both the landline and mobile phone networks, on the performance of FVC using formant trajectories. Trajectories for the first three formants were computed from tokens of the vowel /iau/ of 60 female Chinese speakers, it used both human-supervised and fully-automatic formant tracking systems for features extraction. Measurements were made under different conditions: (i) high-quality (studio recordings), (ii) landline-to-landline, (iii) mobile-to-mobile, and (iv) mobile-to-landline. The speech files were transmitted through an actual mobile phone network, though, the network origin (i.e., GSM or CDMA) and the channel conditions during transmission were not noted in this study. The telephone-quality recordings were treated as offender samples and the high-quality recordings as suspect samples. When only high-quality recordings were used for both the suspect and offender samples, the FVC results using human-supervised formant measurements outperformed the fully-automatic formant tracking system. However, for all the mobile phone speech experiments, both tracking systems showed similar FVC performance and this was worse than the high-quality speech experiments.

2.4 Chapter summary

The Bayesian LR framework is considered by many to be the logically correct way of quantifying the strength of speech evidence and expressing the outcome of an FVC analysis. The LR approach takes into account the prosecution and defence hypotheses and does not involve any binary decisions.

Many different parameters can be used for comparing speech samples and these are broadly classified into auditory and acoustic parameters. The choice to use acoustic parameters in this study was motivated by the fact that they are less subjective and the extraction process is more easily replicable as compared to auditory analysis. Further, a large number of
them can be automatically extracted and they have been shown to outperform the auditory parameters in a number of speech applications.

Several studies have investigated the differences between the landline and mobile phone networks and their varying impacts on the speech signal. However, these studies have focused on the GSM network when studying the impact of mobile phone networks and none (to the author’s knowledge) on the CDMA network. Care must be taken when analyzing mobile phone speech as it could be transmitted over either the CDMA or GSM mobile phone networks. Both technologies incorporate different processing and encoding schemes which could give rise to differences in their impact on FVC analysis. The lack of research studies focusing on the CDMA network and its impact on speech is clear, not only in the FVC arena, but even in speaker and speech recognition applications. It is understandable that the GSM network dominates the market. However, if we look at the world as whole, almost 25% of the world’s population is using the CDMA network, which is something that cannot be ignored. These have been some of the motivations for this research to focus on the CDMA network.
Chapter 3

The Bayesian Likelihood Ratio Framework

This chapter provides background information necessary to understand the likelihood-ratio framework and the various probabilistic models currently used for calculating an LR. The latter sections of this chapter discuss different tools for estimating the performance of an FVC system, such as the Log-likelihood-ratio Cost (Cllr), Tippett plots, Credible Interval (CI) and APE plots.

3.1 Procedures for computing the likelihood ratio

Different probabilistic models have been established for computing the LR. These models generally comprise two major components - similarity and typicality. The numerator of the LR formula calculates the similarity between suspect and offender samples, whereas the denominator estimates how typical they are with respect to a relevant population [53]. Some of the probabilistic models to calculate LR values include: Univariate (UV), Univariate Kernel Density (UKD) [54], Multivariate Kernel Density (MVKD) [10, 15], Gaussian Mixture Model-Universal Background Model (GMM-UBM) [15, 77] and Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [71]. The latter approach is an alternative model to MVKD, which has been developed as part of this work to overcome
computational problems associated with the MVKD. Details of this approach are discussed in Section 3.1.4.

3.1.1 Univariate Kernel Density (UKD)

With UKD, the within-speaker data is modelled using a normal distribution. However, it is generally accepted that the between-speaker statistics cannot be modelled accurately using a normal distribution [11]. So with UKD, the background data is instead modelled by a sum of kernels of equal weights with each being placed over the mean of an individual speaker in the background. The following equations show the mathematical implementation of UKD for computing an LR [54].

\[
LR = K \times \exp \left\{ -\frac{(\bar{x} - \bar{y})^2}{2a^2\sigma^2} \right\} \times \sum_{i=1}^{k} \exp \left\{ -\frac{(m + n)(w - \bar{z}_i)^2}{2(\sigma^2 + (m + n)s_k^2\lambda^2)} \right\}, \tag{3.1}
\]

where

\[
s_k^2 = \sum_{i=1}^{k} \frac{(\bar{z}_i - \bar{z})^2}{k-1} - \hat{\sigma}^2 \frac{k}{k'},
\]

\[
\sigma^2 = \frac{1}{(k+1)} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{N} \frac{(z - \bar{z}_j)^2}{N-1} - \hat{\sigma}^2 \right\},
\]

\[
\hat{\sigma}^2 = \frac{1}{k} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{N} \frac{(z - \bar{z}_j)^2}{N-1} \right\},
\]

\[
K = \frac{k \sqrt{\sigma^2 + ms_k^2\lambda^2}}{a\sigma \sqrt{mn} \sqrt{\sigma^2 + (m + n)s_k^2\lambda^2}} \sqrt{\sigma^2 + ns_k^2\lambda^2},
\]

\[
z = \frac{\bar{x} + \bar{y}}{2},
\]
3.1 Procedures for computing the likelihood ratio

\[ a = \sqrt{\frac{1}{m} + \frac{1}{n}}, \]
\[ w = \frac{m\bar{x} + n\bar{y}}{m + n}, \]

and

\( \bar{x}, \bar{y} \): means of offender and suspect data, respectively;

\( \bar{z} \): mean of an individual speaker in the background;

\( \sigma^2, s_k^2 \): within- and between-speaker variances, respectively;

\( \bar{\sigma}^2 \): suspect and offender combined variance;

\( \lambda \): smoothing factor, its value either being chosen subjectively by comparing the density estimate curve with the actual histogram, or automatically using a number of standard procedures [47];

\( N, k \): number of tokens per speaker, number of speakers in the background, respectively;

\( m, n \): the number of tokens of the offender and suspect data, respectively.

The parameter \( \sigma^2 \) shown in 3.1 is different from the one given in [54] in which \( \sigma^2 \) was originally used to describe the within-offender variance rather than overall within-speaker variance. This modification has been made for the PCAKLR approach in order to align with the way in which this parameter is estimated in MVKD, the overall aim being to ensure functional equivalence between MVKD and PCAKLR (because PCAKLR uses UKD for computing LRs). With respect to the different terms of the UKD formula of (3.1), the first term in the numerator accounts for the similarity between suspect and offender evidence. The second term in the numerator accounts for the location of the combined evidence (i.e., suspect and offender samples) in the overall population distribution, which in turn is a measure of typicality. The first and second terms in the denominator of (3.1) account for the typicality of suspect and offender evidence, respectively.

Since the voice is multidimensional [78], it is possible to calculate individual LRs for a number of speech parameters. If the speech parameters are uncorrelated, the resultant LRs computed using UKD can be combined through simple multiplication to produce an overall LR value (i.e., the naïve Bayes approach [79]). However, the assumption of independence
between speech parameters does not hold well as speech is generated from the same vocal tract and thus a degree of correlation between speech parameters is to be expected. Ignoring such correlations can result in over- or under-estimation of the strength-of-evidence [79, 11], an example of which occurred in the case R v Clarke [11], though this was not a case involving voice comparison. The MVKD algorithm to be discussed next is a widely used approach for taking account of correlation between parameters when the number of parameters is small, the PCAKLR procedure being proposed (see Section 3.1.4) as an alternative when this number is large.

3.1.2 Multivariate Kernel Density (MVKD)

Given that MVKD is well documented in the literature, it is not the intention here to discuss the algorithm in great detail. The intention is to highlight the cause of the lack of robustness of this algorithm when a large number of parameters is used. In the various experiments reported in this thesis, Morrison’s Matlab implementation of the Aitken and Lucy MVKD formula has been used 1.

It is not difficult to identify computational weaknesses in the MVKD algorithm. For example, two matrices, $D_1$ and $D_2$ [10] are used extensively throughout the formulation, these being the covariance matrices for offender and suspect data, respectively. The inverses of these matrices, namely $D_1^{-1}$ and $D_2^{-1}$, are required at a number of stages in the algorithm. There are also instances where these inverted matrices are then inverted again, as, for example, with the term:

$$\left[-\frac{1}{2}(y^* - \bar{x}_1)^T\{(D_1^{-1} + D_2^{-1})^{-1} + h^2C\}^{-1}(y^* - \bar{x}_1)\right].$$

However, in practice, the condition number of either or both of $D_1$ and $D_2$ can be quite large, particularly if the number of input parameters is large, which means that they are ill-conditioned, leading to the possibility of significant inaccuracies with respect to the

3.1 Procedures for computing the likelihood ratio

Chapter 3

computation of their inverses [80, 81, 82]. Another cause of fragility with the MVKD formula is with respect to smoothing of the kernel density estimation. This process becomes a difficult task with the sparseness of data in higher-dimensional spaces. In other words, with practical sample sizes such as those used in the forensics arena, developing a relatively accurate kernel density model is likely to be difficult when using more than about four or five parameters. This phenomenon is commonly referred to as the curse of dimensionality [83]. The fragility of MVKD is investigated in Chapter 4.

3.1.3 Gaussian Mixture Model–Universal Background Model (GMM-UBM)

This approach is commonly used in automatic speaker recognition applications, but has also had success in FVC analysis [77, 84]. The GMM-UBM model differs from the MVKD in two regards. First, the background population (i.e., the UBM) is trained on all the background data pooled across all speakers instead of combining sub-models (or kernels) for each speaker. Second, rather than creating a suspect model from scratch using the often limited amount of speech data available, it is built on the basis of the UBM and uses a procedure called Maximum a Posteriori (MAP) to adapt its covariance matrices, weights and mean values towards a better fit to the suspect data (normally only the mean values of the UBM are adapted). An LR is then calculated as the relative value of two Probability Density Functions (PDFs), the adapted suspect model and the UBM, at each value extracted from the offender samples. Each value provides an LR and then the LR results are combined by taking their product, ignoring the fact that the input parameters are correlated. As mentioned earlier in Chapter 1, this approach typically requires a large amount of data to begin with, which can be problematic in real forensic casework. Speech parameters used for the GMM-UBM training do not usually relate in a meaningful way to words or structures, individual vowel or sounds, but rather they account for fluctuations in the speech envelop as a function of time [15]. The GMM-UBM model was originally designed for data-stream experiments and typically requires a large amount data. Therefore, it is beyond the scope of this study.
3.1 Procedures for computing the likelihood ratio

3.1.4 Principal Component Analysis Kernel Likelihood Ratio (PCAKLR)

This new approach, namely Principal Component Analysis Kernel Density (PCAKLR) [68], for computing LRs presented in this thesis is a combination of two processes. Firstly, the speech parameters selected for FVC analysis are transformed into new sets of uncorrelated parameters. Then an LR value for each of these transformed parameter sets is determined using UKD. Given the assumption of uncorrelated parameter sets, the resulting LRs are then combined by taking their product in accordance with the naïve Bayes approach.

PCA is a widely-used tool in modern data analysis [85], which can be used for extracting information from large correlated data sets. It can also be used to convert observations of correlated variables into new sets of uncorrelated parameters, called principal components, which is the feature being utilised in this PCAKLR technique. These principal components are guaranteed to be independent provided the input data set is jointly normally distributed [85, 86, 87]. However, even if this assumption of normality is not strictly valid for the particular speech parameters used, PCA still produces highly uncorrelated principal components. PCA analysis arranges these principal components according to their significance or importance with respect to representing the information contained in the input data. Specifically, the first principal component has the highest information content, with the last having the least. Given this ordering of principal components according to their information content, one could if one wished reduce the dimensionality of an input data set by discarding one or more of the higher order principal components, a process called dimensionality reduction [88]. However, this has not be done in the case of PCAKLR, the rationale being that the more information used in an FVC, potentially, the higher the strength-of-evidence that will result.

Determining principal components from a set of input data is done as follows. Consider that $N$ observations have been taken of each of $M$ input parameters $\alpha_i, i = 1, 2, \ldots M$. Let the combined set of observations comprising the input data be arranged in a matrix $X^{(M \times N)}$, where each row of $X$ contains $N$ observations of one of the input parameters. The first step in the PCA analysis is to mean-shift the observations for each parameter to zero (i.e., each
3.1 Procedures for computing the likelihood ratio

row of $X$, producing a new matrix $Y^{(M \times N)}$. The next step is to evaluate the covariance matrix, $C^{(M \times M)}$, for $Y$. The diagonal elements of $C$ are then the variances of the observations for each of the input parameters and the off-diagonal elements are the covariances between parameter observations. Since $C$ is square, its eigen vectors $v_i^{(M \times 1)}$, $i = 1, 2, \ldots M$, and eigen values $\lambda_i$, $i = 1, 2, \ldots M$, can be computed, the magnitude of each $\lambda_i$ being a measure of the importance or information content of its associated eigen vector. These eigen vectors are arranged in the matrix $V^{(M \times M)}$, the columns of which are the eigen vectors, $v_i$, ranked left to right in descending order according to their eigen value. The mean-shifted input data matrix, $Y$, now gets transformed into a new data matrix, $Z$, by computing the matrix product $Z^{(M \times N)} = V^T \cdot Y$, where $T$ has its usual meaning of transpose. As was the case with $X$, the matrix $Z$ contains a set of $M$ row vectors each containing $N$ observations of what is now a transformed parameter set $\phi_i$, $i = 1, 2, \ldots M$. This parameter set is entirely different to the original parameter set $\alpha_i$, $i = 1, 2, \ldots M$ associated with $X$. Effectively, the matrix product $V^T \cdot Y$ projects the mean-shifted original input data onto the eigen vector space of the covariance matrix $C$. Specifically, the projection of this data onto the first eigen vector produces the first principal component, $z_1$; the projection onto the second eigen vector the second principal component, $z_2$; and so on to $z_M$. The first principal component, $z_1$, is now the set of $N$ observations of the transformed parameter $\phi_1$. It carries the most information about the original input data contained in $X$, with successive principal components carrying successively less information until the last principal component, $z_M$, which carries the least information. Given that the set of principal components, $z_i$, $i = 1, 2, \ldots M$, not only collectively carries all the information contained in the original input data, but should also be orthogonal, it is an ideal set from which to compute an overall LR value. From the set $z_i$, $i = 1, 2, \ldots M$ a set of LR values, $LR_i$, $i = 1, 2, \ldots M$, is computed using UKD applied to each. Because these resulting LRs have been determined from parameters sets that are orthogonal, an overall LR can be computed by taking their product.

The PCAKLRR approach should be implemented in an actual FVC analysis as follows. For the sake of illustration, consider that the speech data of offender and suspect, denoted as $Off$ and $Sus$ are to be compared against data for a representative background population comprising ten speakers $S_1, S_2, \ldots S_{10}$. Assume now for each speaker, five tokens are available for analysis and that for each token, observations of three acoustic parameters, $\alpha_1$,
3.1 Procedures for computing the likelihood ratio

\( \alpha_2 \) and \( \alpha_3 \), are extracted. Thus in this example the total number of tokens \( N = 60 \) (5x12, the combined tokens for suspect, offender and background population) and the number of parameters/tokens \( M = 3 \). All of this data must be arranged in an input data matrix, \( X^{(3 \times 60)} \) as shown in Figure 3.1. Each row of \( X \) contains observations of one of the acoustic parameters, \( \alpha_i \), and the columns of \( X \) are the observations of \( \alpha_i \) per token. The columns of \( X \) in this example are grouped into blocks of 5 corresponding to token data for each speaker (5 tokens/speaker). The second step is to produce the mean-shifted matrix \( Y^{(3 \times 60)} \) by calculating the mean of each row of \( X \) and subtracting it from the data in that row. The covariance matrix, \( C^{(3 \times 3)} \), is then computed for \( Y \). The eigen values and eigen vectors for this square matrix are determined, from which the matrix \( V^{(3 \times 3)} \) is formed, its columns are the eigen vectors of \( C \) arranged in descending order, from left to right, according to their corresponding eigen values. A set of principal components is generated by computing the matrix product \( Z = V^T Y \), where the rows of \( Z \) are the required principal components arranged in descending order, top to bottom, according to their information content. The first principal component, \( z_1 \), contains 60 observations of the transformed parameter \( \phi_1 \), and similarly for the other principal components. The UKD analysis is then applied to the observations contained in each row of \( Z \) producing three LRs: LR\(_1\), LR\(_2\) and LR\(_3\). The product of these three gives an overall LR value.

An important point needs to be noted from this example. The principal components must be computed from the combined set of data containing observations of suspect, offender and background speech samples. The reason for this is that the PCA analysis is effectively transforming all of the input data into a new vector space. If PCA analysis was applied to suspect, offender and background data separately, each set of input data would be transformed into a different vector space. It would then not be possible to make any meaningful comparisons between the suspect and offender observations as one would not be comparing ‘like with like’. All of the input data must be transformed into the same vector space in order to ensure that each of the subsequent UKD analyses is meaningful.

It needs to be mentioned here that I have developed the PCAKL approach in collaboration with another PhD student, Balamurali B.T. Nair, who also needed this tool to investigate impact of the GSM network on FVC analysis [89, 68]. Balamurali worked on the PCA
Figure 3.1: Block diagram showing the arrangement of data in the PCAKL approach

---

\( \mathbf{X}^{(3 \times 60)} \)

\[ \begin{aligned}
\text{Off} & \quad \text{Sus} & \quad S_1 & \quad S_2 & \quad S_{10} \\
\alpha_1 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\alpha_2 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\alpha_3 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\end{aligned} \]

\( \text{Off} \): Offender  \\
\( \text{Sus} \): Suspect  \\
\( S_i \): Background population

---

\( \mathbf{Y}^{(3 \times 60)} \)

\[ \begin{aligned}
\varphi_1 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\varphi_2 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\varphi_3 & \quad [1 \times 5] & [1 \times 5] & [1 \times 5] & \ldots & [1 \times 5] \\
\end{aligned} \]

---

\( \text{UKD} \rightarrow \text{LR}_1 \rightarrow \text{LR}_2 \rightarrow \text{LR}_3 \)
3.2 Ways of fusing the likelihood ratio results

As previously discussed in Chapter 2, the Bayesian approach allows for combining of LR results, calculated from parameters of the same speech segment, by taking their product on the assumption of zero correlation between these parameters. However, this assumption does not hold well in speech as all parameters are generated from the same vocal tract and therefore a degree of correlation between them is to be expected. As a result, various techniques such as the multivariate and PCA analyses have been proposed to remove the within-segment correlations. Another type of correlation, which has not yet been discussed is between-segment correlation. Forensic speech scientists often extract information from multiple speech segments and thus there is a necessity to combine the overall LR results calculated for each individual segment. The standard statistical procedure for doing this in speech forensics is called logistic regression, which is briefly described in the following section. With PCAKLR, however, the incorporation of PCA into the PCAKLR approach provides an alternative method for combining LR for multiple speech segments. This can be achieved by concatenating the speech parameters extracted from multiple segments into one superset and then transforming the whole lot, using PCA, into a new vector space with all transformed parameters being highly uncorrelated.
3.2 Ways of fusing the likelihood ratio results

3.2.1 Logistic regression

With this method a set of weights, called the fusion parameters, is determined from a large set of comparisons, normally using data from a development set. The origin of each of these comparisons is known *a priori* to be same- or different-speaker. This fusion system requires two or more LRs calculated from different phonemes (or speech segments) for each comparison. The logistic regression is then trained to adapt the fusion weights while minimizing the measure of accuracy (i.e., the Log-likelihood-ratio Cost $C_{llr}$). This is then used to combine LRs from the testing set for multiple segments (note that the development set must be exclusive of the testing set). Further details about this procedure can be found in [90, 91].

3.2.2 PCAKLR with No Fusion (PCAKLR$_{NF}$)

Before discussing how this method can be used for combining LRs from multiple segments, one needs to appreciate first the low degree of correlation between the resulting principal components [87]. For the sake of illustration, the “Spearman” correlations [92] between $\frac{14!}{2!\times(14-2)!} = 91$ pairs of 14 LPCC coefficients were extracted from the NRIPS database. This database comprises 297 Japanese male speakers and includes recordings of tokens of five different vowel phonemes /a i e o u/. The LPCCs were extracted from a 14th-order linear prediction analysis of a 25.6 ms representative segment of a vowel. The resulting LPCCs dataset was determined and provided to this research by [14, 93]. The resulting within-segment correlations are plotted in 3-D in Figure 3.2(i) for the vowel phoneme /a/. The correlation values between the PCA transformed parameters are also shown in 3-D in Figure 3.2(ii).

Given that PCA is an effective tool to extract information from confusing data sets that may well be correlated in some manner, should also be able to handle the correlation between different speech segments. The parameters from any speech segments can be merged to form a superset which is then transformed using PCA to produce a set of uncorrelated parameters. This in fact suggests an alternative way to the conventional approach of combining evidence from different speech segments, namely the Logistic-regression fusion
Figure 3.2: Within-segment correlations for /a/ for (i) LPCCs, (ii) PCA transformed parameters; between-segment correlations for /a i e o u/ for (iii) LPCCs, and (iv) PCA transformed parameters.

[94, 95, 65]. This strategy is referred to in this thesis as PCAKLR$_{\text{NF}}$ (i.e., PCAKLR with no fusion). In order to demonstrate the capability of PCA to take account of between-segment correlations, (e.g., between vowel phonemes), the “Spearman” correlations for a superset of 70 LPCC coefficients have been computed (i.e., five sets of 14 LPCC coefficients extracted from the vowel phonemes /a i e o u/) (plotted in Figure 3.2(iii)). The correlations between their corresponding transformed parameters are shown in Figure 3.2(iv). Figures 3.2(i) and 3.2(iii) show that the within-segment and between-segment correlations, respectively, associated with the LPCCs are quite high (i.e., the off-diagonal peaks and valleys). With respect to the transformed parameters, it is clear from Figures 3.2(ii) and 3.2(iv) that these are far less correlated than the LPCCs. Nonetheless the non-linear correlations are not zero, an aspect which will introduce some error into the subsequent computation of an overall LR value computed using the naïve Bayes approach.
The PCA stage of the PCAKLR approach is capable of handling the within-segment correlation (i.e., within a phoneme in this case) as well as between-segments (i.e., between phonemes in this case) correlation, thus eliminating the subsequent need to fuse the LR results. However, the resulting LRs still require calibration [65, 95, 96]. As will be shown in Chapter 4, this alternative approach of combining LRs for multiple speech segments using PCAKLR, which is referred to as PCAKLR_{NF}, has shown better results than when using logistic regression.

### 3.3 Tools for estimating the performance of an FVC analysis

This section discusses four different numerical and graphical tools used in this study for assessing the impact of various aspects of the CDMA network on the performance of FVC analysis. The tools discussed are: the Log-likelihood-ratio Cost (Cllr) [11, 97, 96, 65, 62], Credible Interval (CI) [98, 63, 62], Tippett plots [65, 62, 60] and APE [99, 97]. Other tools such as the Detection Error Trade-off (DET) and Equal Error Rate (EER) have not been used in this study for the same reasons of not using the classification-error rate metric, as discussed in Section 2.2.3.

#### 3.3.1 Numerical Methods

##### 3.3.1.1 Validity (Log-likelihood-ratio Cost)

The validity or accuracy of an experiment compares the calculated LR values in the test set with knowledge about whether each LR value was the result of a same- or different-origin comparison. Comparing these LR values with the actual output determines whether the estimation is correct or not. One such metric recommended for use in the speech forensics arena is $C_{llr}$ [96, 65, 62]. $C_{llr}$ captures the goodness of a set of LRs derived from a test data set using the following formula.
3.3 Tools for estimating the performance of an FVC analysis

\[ \text{C}_{llr} = \frac{1}{2} \left( \frac{1}{N_{so}} \sum_{i=1}^{N_{so}} \log_2(1 + \frac{1}{N_{soi}}) + \frac{1}{N_{do}} \sum_{j=1}^{N_{do}} \log_2(1 + N_{doj}) \right) \]  

(3.2)

where \( N_{so} \), \( N_{do} \) are the number of same- and different-speaker comparisons, respectively; \( LR_{so} \), \( LR_{do} \) are the LRs determined for same- and different-speaker origins, respectively. \( C_{llr} \) is the average of two parts. The first part of the formula is the mean of a function of LRs obtained from same-origin comparisons, whereas the second part is the mean of another function of all LRs obtained from different-origin comparisons. Ideally, a same-origin comparison should result in a large positive LLR whereas a different-origin comparison should result in a large negative LLR. Therefore, large negative LLRs for same-speaker comparisons and large positive LLRs for different-speaker comparisons are contrary to fact, and are thus heavily penalized in the \( C_{llr} \) formula. Generally \( C_{llr} \) values below 1 indicate that an FVC analysis is providing some useful information. The lower the \( C_{llr} \) value, the more accurate the results of an FVC analysis, and vice versa.

3.3.1.2 Reliability (Credible Interval)

The precision measurement or reliability is an important metric in FVC which quantifies the amount of variation in an LR calculation. The variation is normally a function of changes in the background population and/or source variability [98, 63, 62]. Estimating the reliability by changing the background population is a common practice in forensic DNA profile comparison [100]. This is because DNA profiles do not change over time, so there is no source variability. Unlike DNA, however, variations in a person’s speech can change from one occasion to another. Therefore, the reliability of FVC can be estimated either by changing the background population or using different recording sessions for the same speaker, thereby accounting for source variability. In these experiments, source variability has been used in order to assess reliability of an FVC analysis.

Many routines have been developed to quantify reliability. One of them, proposed in 1998 [80], is based on a statistical technique called bootstrapping and accounts for variability in both the source and background populations. However, in FVC it is common to use the method published in 2011 [62], which accounts for source variability only, and this is
the method used in this thesis. This method proposes the use of credible interval (CI) as a measure of reliability. CI in FVC tries to answer the question: How much variability in estimating the strength-of-evidence is expected due to variability in the measurement of the speech parameters, if the comparison process was repeated several times across different recording sessions?

Once the CI is estimated, one can be confident that the true value of the LR would lie within the 95 percentile of it. Two approaches can be used to estimate the CI: Parametric and non-parametric [62]. The parametric approach is recommended whenever homoscedasticity can be assumed about the LRs distribution, (i.e., the amount of variation in LR values is the same from one comparison to another). For heteroscedastic distributed LRs (i.e., the amount of variation in LR values differs from one comparison to another), the non-parametric CI calculation is appropriate. Heteroscedasticity is commonly the case in FVC, so the non-parametric CI calculation has been adopted here.

In the parametric approach, calculating CI is relatively simple to implement. One would only need to calculate the sample variance of the pooled LLRs irrespective of their underlying origin (i.e., same- or different-speaker origin). The estimated variance is then combined with a t distribution to calculate CI. If homoscedasticity cannot be assumed, then the non-parametric approach is more appropriate. The procedure for calculating CI in this case begins by assuming that the LLRs deviation-from-mean distribution to be symmetrical in a LLR space. In other words, all the negative deviation-from-mean values can be flipped over to positive values. Local linear regression is then applied to find the boundary between the 95% lowest and 5% highest magnitude of deviation-from-mean values [62].

### 3.3.2 Graphical Methods

#### 3.3.2.1 Tippett plots

A Tippett plot is a graphical way of presenting the LLR results of an FVC analysis [12], example of which is shown in Figure 3.3. It represents the cumulative proportion of the
3.3 Tools for estimating the performance of an FVC analysis

Figure 3.3: Tippett plot showing same- and different-speaker comparison results of an experiment.

LLR values. Since large positive LLR values support the same-speaker hypothesis and large negative values the different-speaker hypothesis, the further apart these curves (the same-speaker curve towards the right and the different-speaker curve to the left), the better are the results.

3.3.2.2 APE plots

An Applied Probability of Error (APE) plot [97, 96] is a graphical representation of $C_{llr}$ which teases out the losses of an FVC system under evaluation. Ideally a perfect FVC system should have zero loss (i.e., $C_{llr} = 0$). In reality, though, all systems do have loss, this being the sum of two components, discrimination loss ($C_{llr_{min}}$) and calibration loss ($C_{llr_{cal}}$). $C_{llr_{min}}$ corresponds to the lowest $C_{llr}$ that can be achieved while still preserving discrimination power. This is determined by optimizing $C_{llr}$ using the calibrated LRs of a system under evaluation in an iterative process called Pool Adjacent Violators (PAV) [30, 31]. $C_{llr_{cal}}$ can be obtained by subtracting $C_{llr_{min}}$ from $C_{llr}$.
3.3 Tools for estimating the performance of an FVC analysis

An APE-plot comprises a number of APE-curves and bar graphs, examples of which can be found in Chapters 6 to 10 (e.g., see Figure 3.4). An APE-curve, which plots error-rate against logit prior, comprises three curves: green, red and black. The red curves show the error-rate of the actual LLRs of the system under evaluation. The height of the red portion of the bar graph represents $C_{llr_{cal}}$, which is proportional to the area between the green and red APE-curves. The green curve shows the error-rate that could have been obtained if LLRs provided under evaluation were perfectly optimized. The value of $C_{llr_{min}}$ is reflected in the height of the green portion of the bar graph and is proportional to the area under the green APE-curve. Finally the black curve shows the error rate of the reference system (i.e., $C_{llr} = 1$ or LLRs = 0).

Figure 3.4: APE plots showing the losses in $C_{llr}$ from two different FVC experiments.
### 3.4 Chapter summary

The LR framework is based on determining how much more likely it is to observe differences between the offender and suspect speech samples assuming they are uttered by the same speaker rather than being uttered by different speakers. The most widely used probabilistic models for calculating LRs are the Multivariate Kernel Density (MVKD) and Gaussian Mixture Model–Universal Background Model (GMM-UBM). Comparisons have been previously reported with respect to the GMM-UBM and MVKD and have shown that both systems have a similar performance when investigating individual phonemes. However, the GMM-UBM is reported to generally outperform MVKD analysis upon fusing the LR results. The downside of GMM-UBM is that a large amount of data typically required to construct the GMM and UBM models.

In this study, a new procedure is presented for calculating an LR, namely the Principal Component Analysis Kernel Density (PCAKLR). As with MVKD analysis, background data is modelled using a set of kernels positioned around the mean value of each speaker. As will be shown in the next chapter, PCAKLR has a comparable performance to the MVKD, yet it is computationally robust irrespective of the number of input parameters used. Using MVKD for a large number of parameters, greater than the number it was originally designed for, can be problematic and may result in ill-conditioned matrices. This can lead to significant errors in the computation of LR values. The PCAKLR approach also provides an alternative procedure to logistic regression for combining LR results from multiple speech segments. The speech parameters from all speech segments can be grouped into one superset and then it is left to the PCA stage of PCAKLR to remove the between- and within-segment correlations. The resulting LLRs from all transformed parameters are then added and calibrated.

A number of tools and metrics can be used to estimate the performance of an FVC analysis, these are: the Log-likelihood-ratio Cost ($C_{llr}$), Credible Interval ($CI$), Tippett plots and APE plots. The Log-likelihood-ratio Cost ($C_{llr}$) is a more appropriate metric for estimating the accuracy within the LR framework. Previous metrics such as the classification-error rate and correct-classification rate constitute a binary decision and do not align with the
principles of the Bayesian approach. The Credible Interval (CI) is also an appropriate measure of precision for the LR framework. CI is normally used in FVC to estimate variations across different recording sessions. However, it can also be used to measure variations with respect to any other factors, such as different background populations or different recording qualities, etc. Tippett plots are used to visually inspect the LR results, where these show the proportion of LRs higher or lower than a certain LLR. The further the curves are apart, the better the discrimination between speakers. APE plots are used to tease out information about the cause of loss in $C_{llr}$, which can be either a discrimination or calibration loss. Very low calibration loss indicates that the system used for FVC analysis is naturally calibrated. Discrimination losses relate to the discrimination power of the input parameters and the ability of the system to use those parameters efficiently.
Chapter 4

Comparative Performance of MVKD vs. PCAKLR

4.1 Experimental methodology

The accuracy and robustness of the PCAKLR technique was assessed and compared with the MVKD by undertaking two sets of experiments [68]. The first of these used speech acquired from the Timit database [24]; the second used a database acquired by NRIPS for forensic speaker recognition tests [25]. The first experiment was used solely to examine the robustness of PCAKLR under circumstances where the MVKD algorithm exhibits its fragility. The second experiment, with the NRIPS database, was used to investigate and compare the performance of PCAKLR with MVKD when applied to more real-world forensic situations.

The TIMIT speech data is sampled at 16 kHz and coded into 16 bits. Since the purpose of this experiment was to examine the robustness of PCAKLR in high dimensional spaces where the available data is sparse, a small number of speakers has been used here (eleven speakers). The speakers chosen are males speaking the northern American dialect (the DR2 subset). A set of 15 LPCCs have been extracted from tokens of the speech segment /æ/. This was extracted from words embedded in connected and read speech. For a particular experiment, one of the eleven speakers was chosen to be the suspect, one the offender, and
the remaining speakers becoming the relevant background population. This experimental setup allows 55 different-origin comparisons and 11 same-origin comparisons to be conducted. Separate sub-experiments were then conducted using different numbers of LPCCs in the case of MVKD and transformed parameters in the case of PCAKLR (up to a maximum of 15). For example, in the case of MVKD, the first in a sequence of experiments would use only $c_1$, the second $\{c_1, c_2\}$, the third $\{c_1, c_2, c_3\}$ and so on until all 15 cepstral coefficients were being used for a comparison. In the case of PCAKLR, the same procedure was adopted, but used the transformed parameters $\phi_k$. Separate LR values were computed for each sub-experiment for both the MVKD and PCAKLR approaches and then the accuracy of these experiments, in terms of $C_{lir}$, plotted as a function of the number of parameters used.

The NRIPS database [88] comprises non-contemporaneous landline telephone recordings of 297 male Japanese speakers, all of whom were members of the Japanese Police force. The recordings were digitised at 10 kHz using 12-bit quantisation. Two non-contemporaneous recordings of read speech, separated by three to four months, were made for each speaker. The recordings were typically of 70-80 seconds duration and comprised both single- and many-word utterances. They also included tokens of the five standard Japanese vowel phonemes /a i e o u/ and it was these that were used in these experiments. In each recording all of the data was repeated once, thus giving two replicates of each vowel per non-contemporaneous recording session.

In this thesis, the performance of PCAKLR has not been compared to the GMM-UBM model for two reasons. First, the GMM-UBM has been primarily designed for data-stream-based analysis scenarios, whereas MVKD and PCAKLR are primarily designed for token-based analysis scenario. The difference between MVKD and PCAKLR is principally with respect to the number of parameters that can be handled, this being three to four in the case of MVKD, and much larger than this in the case of PCAKLR. Given that vowel tokens have been used in the experiments of this study, it was not clear how PCAKLR can be compared to a data-stream model such as the GMM-UBM model. Second, the main focus of this study was not to compare the performance of FVC systems, but rather the impact of mobile phone network on FVC analysis. The MVKD algorithm was originally intended for use
in this study, but found it to produce inaccurate results when used with a large number of parameters. That was the motivation for developing PCAKL.

4.2 Computational robustness of MVKD vs. PCAKL in high-dimensional spaces

Using both the MVKD or PCAKL analysis procedures for computing LR values, results have been produced for the 11 same-speaker and 55 different-speaker comparisons for the vowel /æ/ using different numbers of input parameters. The accuracy of the resulting comparisons has then been determined using $C_{llr}$ and this plotted as a function of the number of input parameters from 1 to 15 (as shown in Figure 4.1). The vertical scale in Figure 4.1 has been truncated to $C_{llr} = 5$ to aid in comparing the results. Values in excess of this boundary, which occurred for input parameters 7, have not been plotted.

Considering first the results for MVKD, a number of important observations can be made. Firstly, as the number of input parameters is increased from one to six, the $C_{llr}$ steadily decreases, indicating an improving accuracy of the FVC analysis. At seven input parameters, however, there is an unexpected increase in $C_{llr}$, followed by a decrease at eight and nine input parameters, but $C_{llr}$ still unexpectedly high. For 10 input parameters it falls below 0.1, but it is again unexpectedly high for 11 input parameters. For 12 input parameters, $C_{llr}$ falls again to a reasonably low value. An abrupt jump again occurs for 13 input parameters, followed by a very significant drop for 14 input parameters to a value again well below 0.1. It is important to mention that the results in Figure 4.1 for MVKD have not been calibrated, the rationale being that calibration would likely have masked this behaviour.

One is led to question whether the MVKD results of Figure 4.1 are indeed correct for all numbers of input parameters, as such behaviour does not align with expectation. Intuitively one expects that in general the accuracy of an FVC experiment should improve as more relevant information is used in the analysis. This means that one might expect $C_{llr}$ values to decrease as more cepstral coefficients are included in an analysis.
For MVKD this expected $C_{llr}$ trend does indeed seem to be evident for numbers of input parameters in the range one to six. Above this range a consistently improving accuracy no longer occurs. It might be argued that this is because the LPCCs above six not only contained no useful speaker-discriminating information, but actually added in noise to the analysis. But if that is the case, why then should including 10 input parameters, or indeed 14 input parameters, make the situation significantly better.

Before drawing any firm conclusions about these results for MVKD, it is informative to examine the corresponding results for PCAKLR which are shown in the same figure. Here there is indeed a relatively smooth relationship between $C_{llr}$ and number of input parameters which align with the expected trend.

Figure 4.1: $C_{llr}$ vs. number of input parameters for MVKD (red) and PCAKLR (green) for the speech segment /æ/.
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

A comparison of the LR results of MVKD and PCAKLR in Figure 4.1 give further justification for questioning the accuracy of the MVKD analysis for all numbers of input parameters. Matrices of the MVKD algorithm which required inversion have been examined in the case of 7, 8, 9, 11 and 13 input parameters and were found to have very high condition numbers ($\geq 10^6$), which meant that they were ill-conditioned and thus significant errors in the inversion process were likely to have occurred. Conversely, for other numbers of input parameters the matrix condition numbers were typically quite low ($\leq 10^3$). Thus it can be concluded that, for this set of experiments, the MVKD $C_{llr}$ results for 7, 8, 9, 11 and 13 input parameters are incorrect, and this occurred due to problems associated with the inversion of ill-conditioned matrices.

Given that the MVKD algorithm was not designed for numbers of parameters in excess of about three or four, it is perhaps not surprising that it sometimes fails when the number of parameters exceeds this. This fragility is linked to matrix inversion and the difficulty of smoothing the kernel estimation in more than about four dimensions. It is true, as demonstrated by the results in Figure 4.1, that MVKD can sometimes produce reasonable results for input parameter numbers above 4, but from this experiment it can be concluded that it cannot always be guaranteed to do so. Fortunately PCAKLR does not suffer from such fragility issues and can be used for quite large numbers of input parameters even where the number of speakers involved is small.

4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

4.3.1 MVKD vs. PCAKLR for four parameters

The purpose of this experiment, using the NRIPS-database, was to compare performance of the MVKD and PCAKLR for a small number of input parameters, the rationale being that MVKD can be reasonably expected to produce correct results in this situation, thus ensuring a fair comparison between algorithms. Using the testing data set comprising 99 speakers, FVC results have been computed for 4,851 different-speaker and 99 same-speaker
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

comparisons for each of the vowels /a i e o u/ using four LPCCs as input for MVKD and the corresponding four transformed parameters as input for PCAKLR. The results for individual vowels have then been fused using logistic regression for both MVKD and PCAKLR and plotted using Tippett plots (Figure 4.2). The accuracy of the resulting comparisons has also been computed using $C_{llr}$, this parameter being examined further in terms of APE plots (Figure 4.3).

Figure 4.2: Tippett plot showing performance of PCAKLR vs. MVKD based on four speech parameters using conventional fusion.
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

Chapter 4

Figure 4.3: APE-plot showing the $C_{llr}$ discrimination/calibration performance of PCAKLR vs. MVKD based on four speech parameters using conventional fusion.

Considering first the accuracies of the resulting FVC analyses, for MVKD $C_{llr} = 0.374$ and for PCAKLR $C_{llr} = 0.389$. These values are very close, though MVKD has produced a marginally more accurate result. An examination of the resulting Tippett plots shows that in terms of same-speaker comparisons the two routines have produced virtually identical results. However, there is a slight difference in their different-speaker performance, with MVKD again giving marginally better performance.

The results have also been examined using APE plots. As discussed in Chapter 3, these plots are used in FVC to tease out information about the losses in $C_{llr}$ for the system under evaluation. Examination of $C_{llr}$ using the APE-plot shows that MVKD has marginally outperformed PCAKLR in terms of both discrimination loss (MVKD $C_{llr_{min}} = 0.330$, PCAKLR $C_{llr_{min}} = 0.342$), and calibration loss (MVKD $C_{llr_{cal}} = 0.044$, PCAKLR ($C_{llr_{cal}} = 0.047$). Overall, though, it can be concluded from this experiment that PCAKLR performs quite similarly to MVKD for small numbers of parameters.

4.3.2 MVKD vs. PCAKLR$_{NF}$ for four parameters

As previously mentioned, PCAKLR has the ability to handle simultaneously not only correlation between input speech parameters, but also correlation between speech segments,
thus eliminating the need for logistic-regression fusion in order to combine results for different vowels. It is important to note, though, that results produced using PCAKLRF, as the process is referred to, still need to be calibrated using a development set. The third set of experiments using the NRIPS database investigates and compares the performance of PCAKLRF with MVKD and PCAKL. Figure 4.4 shows the combined results for all five vowels (four LPCCs per vowel) for MVKD (results combined using logistic-regression fusion) and PCAKLRF for 20 input parameters (i.e., five vowels of four LPCCs each are combined in one superset) (Note: the results in this figure for MVKD are the same as those presented in Figure 4.2). As before, for MVKD $C_{llr} = 0.374$, while for PCAKLRF $C_{llr} = 0.359$. The important point to note here is that PCAKL, when implemented without fusion, has now produced a marginally more accurate FVC result than MVKD. Note that the results for PCAKLRF have been calibrated, as have those for MVKD (calibration is an integral part of logistic-regression fusion). Figure 4.5 shows a zoomed in section of Figure 4.4 around the important $LLR= 0$ decision boundary. Arguably, for an FVC, it is the performance in the vicinity of this decision boundary which is more important than the performance for higher magnitude LLRs. Indeed, $C_{llr}$ gives greater weight to performance for small LR magnitude values than larger ones.
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

As can be seen from Figure 4.5, PCAKLR\textsubscript{NF} has outperformed MVKD in this region for different-speaker comparisons, though MVKD is very slightly better than PCAKLR\textsubscript{NF} for same-speaker comparisons. The proportion of different-speakers classified as same-speakers was lower for PCAKLR\textsubscript{NF}, but the proportion of same-speaker misclassifications was almost the same for both MVKD and PCAKLR\textsubscript{NF}.

Figure 4.4: The performance of PCAKLR\textsubscript{NF} vs. MVKD using 4 speech parameters.
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

Figure 4.5: The performance of PCAKLR\textsubscript{NF} vs. MVKD using 4 speech parameters (zoomed version).

For larger LLR magnitudes (see Figure 4.4), PCAKLR\textsubscript{NF} has marginally outperformed MVKD for same-speaker comparisons, but the opposite is true for different-speaker comparisons. The APE-plot in Figure 4.6 shows that the slight improvement in $C_{llr}$ when using PCAKLR\textsubscript{NF} is attributable to the improved discrimination of this approach. It is worth mentioning that $C_{llr_{cal}}$ for both PCAKLR (Figure 4.3) and PCAKLR\textsubscript{NF} indicate a similar calibration performance when used with a small number of parameters.

### 4.3.3 PCAKLR vs. PCAKLR\textsubscript{NF} for a large number of parameters

Figure 4.7 compares the accuracy of PCAKLR with PCAKLR\textsubscript{NF} as a function of the number of input parameters, this being varied from 4 to 14. In all cases PCAKLR\textsubscript{NF} has produced a lower $C_{llr}$ value than PCAKLR, though in a few cases (8 and 11 input parameters) this difference is very small.

The results of Figure 4.7 would seem to provide considerable validation for the PCAKLR\textsubscript{NF} approach to combining results for different speech segments (in this case, vowels) at the
4.3 Performance of MVKD vs. PCAKLR for a small number of parameters

Figure 4.6: The $C_{llr}$ discrimination/calibration performance of $\text{PCAKLR}_{\text{NF}}$ vs. MVKD using 4 speech parameters.

PCA stage of PCAKLR rather than subsequently for individual vowels using logistic regression. It is important to note another aspect of the results of Figure 4.7 for $\text{PCAKLR}_{\text{NF}}$. Take for example the case of 14 input parameters. In reality, the $\text{PCAKLR}_{\text{NF}}$ routine is then taking as input a total of 70 parameters (14 input parameters/vowel $\times$ 5 vowels). The fact that the resulting $C_{llr}$ is lower than in the case of the PCAKLR routine running separately for each vowel and results fused using logistic regression, suggests that there are no mathematical stability issues with the $\text{PCAKLR}_{\text{NF}}$ routine. Indeed, as far as mathematical stability issues are concerned, there appear to be no limit to the number of input parameters that PCAKLR can handle. This is in stark contrast to MVKD that was only ever designed for three to four input parameters and cannot be guaranteed to produce accurate results much above this number.
4.4 Chapter summary

Figure 4.7: $C_{llr}$ vs. number of input parameters for PCAKLRF (red dots) and PCAKL (blue dots).

4.4 Chapter summary

The MVKD algorithm cannot be guaranteed to produce accurate results when used in a high-dimensional space (more than four or five parameters). The situation is even more problematic given the sparsity of data that is typically available for forensic purposes. The PCAKL approach has been found to produce comparable results to the MVKD and can actually outperform it if the logistic regression fusion is bypassed with PCAKLRF. PCAKLRF has consistently produced better results than PCAKL.
Chapter 5

Impact of the CDMA Network on Speech Signals

5.1 Differences between communication technologies in respect of handling the speech signal

5.1.1 Landline vs. mobile phone networks

When undertaking FVC analysis, it is important that forensic scientists are aware that landline phone networks are fundamentally different to the mobile phone technologies currently in use today. Not only this, within the mobile phone arena, there are a number of network providers utilizing different systems, such as Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). These two systems are fundamentally different in their design and internal operation and thus in the way they impact on the speech signal, usually negatively from the perspective of forensic analysis.

One of the key differences between landline and mobile phone networks is in respect of the medium of transmission (or channel). In a landline network, this typically comprises a copper wire that connects the subscriber to his network’s local exchange and then a fibre optic cable carrying the speech data across the network. The properties of these physical
Channels typically do not change with time. The landline channel has a relatively high data carrying capacity, with the speech signal being transmitted at bit rates of 64 kbps. Though the speech bandwidth is limited to approximately 300 – 3400 Hz, the speech signal is still transmitted with a relatively high fidelity. On the other hand, subscribers in a mobile phone network are connected via a wireless channel. This channel is generally poor and its properties can change significantly, over short periods of time (as short as 20 ms), especially when the subscriber is not stationary. This is because the wireless channel is susceptible to many external factors such as: fading, phase distortion and interference from other wireless channels. As a result, the data carrying capacity of these networks is relatively low [1, 3].

Highly innovative strategies have been developed for mobile phone networks to maintain a good quality of service despite these challenges. In order to overcome the low data-carrying capacity of the wireless channel, the speech coding bit rates have had to be reduced from the 64 kbps typically used for landline networks. Consequently this has caused a reduction in speech fidelity. Accommodating the time-varying properties of the channel has necessitated a dynamic change in the speech coding bit rate, while at the same time ensuring an acceptable quality of service for each user [101].

5.1.2 The GSM vs. CDMA network

One of the fundamental differences in the way the GSM and CDMA mobile networks handle the speech signal is linked to their different ways of giving multiple users access to the network. The GSM network uses a combination of time division multiple access (TDMA) and frequency division multiple access (FDMA), whereas the CDMA network uses code division multiple access (CDMA). With FDMA, the frequency spectrum allocated to the network provider is divided into small segments called frequency channels and individual users are allocated to one of these. In the case of TDMA, several users share the same frequency channel, but each is allocated its own time slot in a sequential manner [102]. A major challenge in the design of these mobile networks is the significantly changing quality of these frequency channels over even short periods of time. The rationale underpinning Dynamic Rate Coding (DRC) is to change the source coding bit rate according to changing channel quality in order to achieve different levels of protection for the coded speech signal.
5.1 Differences between communication technologies in respect of handling the speech signal

Chapter 5

Under bad channel conditions a large part of a user’s bandwidth is allocated for protection, which in turn results in a smaller bandwidth being available for the coded speech signal. The opposite happens when channel conditions are good [103, 104, 3]. This process is controlled by the network as a whole, but it is the codec which actually implements it. In the CDMA network, users are given access to the entire spectrum at the same time, thus sharing the same frequency and time resources. However, each user is allocated a unique code that can be identified by the network to separate out the users’ transmissions. Unlike the GSM network, with the CDMA network, the requirement for protection is linked to interference from the number of simultaneous users accessing the system, not poor channel quality. When the number of users is high, more bandwidth is allocated for protection from interference and less for transmitting the speech signal, and vice versa [105, 106].

It is the task of the speech codecs in mobile phone networks to implement the processes described above. They do this by compressing the speech signal to produce a certain source bit rate stream, and then appending to this a number of error protection bits. This is done by first segmenting the speech signal into frames. These are analyzed and bits associated with the speech model parameters transmitted, along with protection bits. At the receiving end, the speech model parameters are used to synthesize the speech signal, the protection bits being used to correct, if possible, any errors arising during transmission [3].

5.1.3 Strategies for assessing the impact of mobile phone networks on speech

Clearly the first task for any study that involves examining the impact of mobile phone networks on speech is collecting a database of mobile phone speech recordings that is representative of the very large number of scenarios that could occur during transmission. A common approach to achieving this to transmit speech between two devices across an actual mobile phone network. Though this might seem an obvious and appropriate strategy, it is in fact very time consuming and far from comprehensive as it is likely to encompass only a small subset of all possible transmission scenarios. The impact of a mobile phone network on the speech signal is dependent on many highly variable factors such as wireless
channel conditions present, capacity (i.e., number of users accessing the system), interference levels, etc. These factors are both time variant and location dependent. Hence, in order to collect a truly representative mobile phone speech database using this approach, it would be necessary to conduct a large number of experiments at different times of the day and in many different locations, for example, city centers (typically referred to as urban canyons), urban and rural areas. Even then, though, there would be no way of knowing whether all possible transmission scenarios had been represented because such information is not available in the received speech signal. Further, experiments would need to be repeated using both the GSM and CDMA network technologies as their underlying design is fundamentally different, leading to differences in the way they impact upon the speech signal.

In this thesis, an alternative approach is proposed for simulating the impact of mobile phone networks on speech, which has the potential to encompass and sample the space of all possible transmission scenarios in a sensible manner. Yet, this approach is less time consuming and makes it possible to relatively easily transform any existing speech database into a mobile phone speech database. This approach is based on the fact that it is only the speech codec in these networks that directly impact the speech signal, and thus any changes that might occur to it [4]. Of course one cannot separate the operation of the codec from what is happening in the network. These codecs have many modes of operation, but switching between these modes is a process initiated either by the network as a whole, in response to changing channel conditions, or by the user’s mobile phone in response to changing speech characteristics. Constraints with respect to how these modes can be changed at any instant in time have been built into these codecs, as well as into the design of the networks as a whole. So, though the number of possible codec operational modes is large, it is possible to group these into a smaller number of representative scenarios.

This approach also makes it possible to investigate certain aspects in isolation from others, whereas with the conventional approach this would not be possible. For example, the impact of DRC can be investigated in isolation from FL and BN (see Chapter 7). The impact of FL can also be studied in isolation from BN at the transmitting end, but the process of DRC can never be completely disabled in EVRC, though. This is because speech frames are
always encoded at different bit rates with EVRC depending on the classification of speech frames (e.g., voiced, unvoiced and transient) as will be discussed next. Nevertheless, it is possible to limit to some extent the change in DRC by using fixed Anchor Operating Points (AOPs) (i.e., OP0, OP1 or OP2) during the coding process (further details about this are discussed in Chapter 8). BN at the transmitting end can also be investigated in isolation from FL, where the change in DRC can again be restricted, but not completely disabled for the same reasons previously mentioned (see Chapter 9).

The most widely used speech codecs in the mobile phone networks are the Adaptive Multi Rate Codec (AMR) in the GSM network [107] and the Enhanced Variable Rate Codec (EVRC) in the CDMA network [5]. Due to time limitations, this study has only focused on aspects of the CDMA network. The GSM network and its impact on speech is currently being examined by another PhD student, Balamurali B.T. Nair. Nair and I have recently published a number of papers which compare influence of the GSM and CDMA networks on FVC analysis [1, 2, 28, 108, 109].

Based on a thorough understanding of various aspects in the CDMA network that can potentially impact the speech signal, I developed a software platform to simulate realistic scenarios of transmission in the CDMA network. I made use of the publicly available routines of the EVRC codec. These are the ones tested and verified by the network operator and implemented in our mobile phones. However, this platform allows access to various features of the codec and designed to operate the EVRC in a manner similar to an actual CDMA network. I have also slightly modified the original codec routines to allow access for other features such as the FL mechanism\(^1\) (details of this platform are discussed in Appendix B).

5.1.4 General description of the EVRC

The EVRC comes in narrowband (NB) and wideband (WB) versions. In terms of their bandpass frequency responses, their low end is at approximately 200 Hz. The upper end of their response is at approximately 3.4 kHz in the narrow band version and extends to 7 kHz.\(^1\) The modifications do not change or affect in any way the functionality of the codec, but rather enabling certain flags in the codec to activate FL.

\(^1\) The modifications do not change or affect in any way the functionality of the codec, but rather enabling certain flags in the codec to activate FL.
kH{}z in the WB version. The wideband version uses two different coding schemes for the signal components at the low frequency (LF) (0-4 kHz) and high frequency (HF) (3.5-7 kHz) bands [110]. The coding model used in the LF region is quite similar to the narrowband EVRC. The coding scheme used for the HF band is based on Linear Prediction Coding (LPC). Though the quality of speech is better in the WB EVRC when compared to the NB version, the latter is still by far the most commonly used currently. Therefore, the discussion here refers only to the NB version. In a narrowband EVRC, the speech signal is sampled at 8 kHz, 16-bit coded and then segmented into frames of 20 ms which can be coded at four output bit rates: 0.8, 2, 4 and 8.55 kbps [5]. With CDMA, the network controls the interference between users by specifying a target Average Data Rate (ADR) over a set of speech frames. The ADR can take on any value between 4 kbps to 8.55 kbps. The EVRC takes the target ADR and maps it into one of three anchor Operating Points (OP): OP0, OP1 and OP2. It is then left to the codec to decide upon a specific bit rate and select an appropriate speech coding technique for each individual frame to achieve the target ADR. It does this by analyzing the speech characteristics of a particular frame. The EVRC incorporates a number of speech coding techniques such as Code Excited Linear Prediction (CELP), Noise Excited Linear Prediction (NELP) and Pitch Period Prototype (PPP), the underlying rationale being to try and ensure reasonably good quality of coded speech irrespective of bit rate. Figure 5.1 shows the bit rate selection mechanism for different anchor operating points in the EVRC.

Each individual speech frame is broadly classified into one of three categories: end of speech, speech and silence. Irrespective of the anchor operating point selected, silence frames are always encoded using a special silence encoder at 0.8 kbps, and the end of speech is encoded at 4 kbps CELP. If a frame is classified as speech, the bit rate selection depends on the selected anchor operating point. If OP0 is chosen, the frame will be encoded at 8.55 kbps CELP. If either OP1 or OP2 are selected, the active speech is further classified as voiced, unvoiced or transition. The unvoiced and transition classifications are encoded with 2 kbps NELP and 8.55 kbps CELP, respectively. Voiced speech frames are grouped into pattern sets of three frames. If OP1 is chosen, the three-frame pattern set is
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

As mentioned earlier in Chapter 1, there are four major aspects of the CDMA network which ultimately impact upon the speech signal, namely (i) Dynamic Rate Coding (DRC), (ii) Frame Loss (FL), (iii) Background Noise (BN), and (iv) Silence Frames (SF). The following discussion provides detailed understanding of the first three aspects and how they are linked to what is happening in the network as a whole. The latter aspect is briefly
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

discussed in this chapter, but its respective impact on the speech signal has not been investigated in this study. This is because in real FVC scenarios, only active speech is used in the comparison process.

5.2.1 Dynamic Rate Coding (DRC)

EVRC has the ability to operate at any arbitrary average data rate (ADR) by adjusting the fraction of voiced and unvoiced frames encoded at 2 kbps (Figure 5.2). This adjustment is done dynamically by comparing the actual ADR over the last 600 active speech frames (i.e., 12 sec of speech) with the target ADR. The adjustment process computes the fraction of 2 kbps frames that can be encoded at 8.55 kbps [111]. The ADR value is then updated every frame using the following formula (Equation 5.1):

\[
ADR = \frac{9600N_1 + 4800N_2 + 2400N_3}{N_1 + N_2 + N_3},
\]

where \(N_1\), \(N_2\) and \(N_3\) are the number of frames that were encoded at 9.6, 4.8 and 2.4 kbps, respectively, over the past 600 active speech frames. The bit rates shown in the above formula are the traffic coding bit rates which are a combination of speech coding bits and error protection bits. The mapping between speech coding bit rate and traffic coding bit rate is shown in Table 5.1.

<table>
<thead>
<tr>
<th>Speech coding bit rate</th>
<th>Traffic coding bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.55 kbps</td>
<td>9.6 kbps</td>
</tr>
<tr>
<td>4 kbps</td>
<td>4.8 kbps</td>
</tr>
<tr>
<td>2 kbps</td>
<td>2.4 kbps</td>
</tr>
<tr>
<td>0.8 kbps</td>
<td>1.2 kbps</td>
</tr>
</tbody>
</table>

The point to note here is that the EVRC uses an internal selection mechanism that constantly changes the bit rate irrespective of commands from the network as a whole. If the network instructs the codec to operate at a particular ADR, the ADR will be first mapped into one of the three anchor operating points as shown in Figure 5.2. If the target ADR
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.2: EVRC mapping of target ADR into different anchor operating points.

chosen by the network is greater than 7.5 kbps, OP0 is selected; if it is between 6.6 kbps and 7.5 kbps, OP1 will be selected; if it is below 6.6 kbps, OP2 is selected. Once the codec decides upon an anchor operating point, it will change a number of the 2 kbps frames into 8.55 kbps frames.

With the CDMA network, a bit rate change can occur as frequently as every frame (i.e., every 20 ms). Further, at each frame, the EVRC can switch between OP1, OP2, and vice versa. As shown in Figure 5.3, there are two possibilities for switching between OP1 and OP2, referred to here as Case (a) and Case (b), but only one for switching back again, referred to as Case (c). The speech frames shown in this example are voiced. It is assumed that the codec is instructed to change the anchor operating point at the time of encoding the second speech frame. In all cases the three-frame pattern resets to begin with 8.55 kbps. Upon resetting, the two subsequent frames are encoded in accordance with the new operating point being selected. Case (b) is a special case known as ’dimming’ which happens under high capacity conditions, where the target ADR is set to 4.8 kbps. With dimming, no speech frame is encoded at a bit rate above 4 kbps apart from the first frame upon switching.
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.3: An example showing the switching mechanism between OP1, OP2 and vice versa in the case of coding voiced speech frames.

In terms of resulting speech quality, there is in fact a continuum of possibilities, which range from very poor to very high-quality speech. Nonetheless, it may be argued that these can be classified into one of three categories: low, medium and high speech qualities corresponding to high, medium and low capacities, respectively. This simplification is motivated by aspects of the codec operation to be discussed next.

The ADR can take on any value between 4.8 kbps to 9.6 kbps. Low capacity conditions (i.e., the number of users present in a cell site is small), result in ADR values in the upper-half of this range (see Figure 5.2) and this is achieved by switching between OP0 and OP1 in some predetermined manner. Such a scenario will produce a relatively high-quality speech. When the network is congested with users (i.e., high capacity), the ADR values are selected from the lower-half of this range, achieved by switching between OP1 and OP2. This in turn produces low-quality speech. Under medium capacity conditions (i.e., the cellular site is neither congested nor has a small number of users), ADR values can be selected from the entire range of possibilities. This is achieved by switching between OP0, OP1 and OP2 in some pre-determined manner, resulting in medium-quality speech.

To illustrate the differences between the different anchor operating points with respect to handling the speech signal, a set of time waveforms and their corresponding spectrograms
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.4: A set of time waveforms showing the impact of coding on the word “eight”. (a) Clean and uncoded speech, (b) speech coded at OP0, (c) speech coded at OP1, (d) speech coded at OP2.

have been produced for a speech file containing the word “eight” and these are shown in Figures 5.4 and 5.5, respectively. Figure 5.4(a) shows the time waveform of the original speech segment. Figures 5.4(b), 5.4(c) and 5.4(d) show the same speech signal after having been coded at OP0, OP1 and OP2, respectively.

Examination of both Figures 5.4 and 5.5 shows noticeable differences between the resulting coded speech signals and the original, uncoded speech. Further, the number of pitch periods for OP0 in Figure 5.4(b) is clearly different from those of OP1 and OP2. Whether these differences might contribute to an improvement or worsening performance in an FVC analysis is something that will be thoroughly investigated in Chapter 7.

The key point to note with the DRC process is that successive 20 ms frames of speech could be coded at significantly different bit rates, which translates into significantly different coding quality. As far as can be ascertained, there has been no thorough investigation thus far on the impact of this DRC process on acoustic parameters important to forensic speech science.
Figure 5.5: A set of spectrograms showing the impact of coding on the word “eight”. (a) Clean and uncoded speech, (b) speech coded at OP0, (c) speech coded at OP1, (d) speech coded at OP2.

### 5.2.2 Frame Loss (FL)

In mobile phone networks, the wireless channel can often be quite poor, increasing the likelihood of frames being lost or corrupted during transmission. In order to maintain good quality speech under such conditions, innovative techniques have been implemented in the speech codecs of these networks. Certain bits are checked by the speech codec to determine whether errors in transmission have occurred. If they have, there is an attempt to correct them using, for example, convolutional coding [7]. If correction is not possible, or a frame is lost in its entirety, the mechanism of lost frames is put in place\(^2\). This broadly involves replacing lost speech data using information from previous received speech data. If successive frames are lost, the codec will continue replacing those, while at the same time decreasing the output level gradually until eventually silence results, a process called

\(^2\)Note: for simplicity here, the term ‘lost frame’ is used to refer to frames that are either lost or irrecoverably corrupted, because in both cases the corrective measure and resulting outcome are the same.
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

This muting process involves repetition of gain parameters with a certain decaying factor. A maximum of 16 successive frames could be replaced in this manner before silence results.

From the perspective of an FVC, this process of handling lost frames by replacing them with speech data derived from the past is clearly of concern, unless those speech sections affected can be detected a priori and excluded from the analysis. However, the EVRC has been designed with speech quality in mind and therefore sophisticated strategies have been put in place to minimize or even eliminate any perceptual artefacts that might arise from the lost-frame process. The result is that subsequent detection of this process from the received speech signal is likely to be very difficult, if not impossible. These strategies include smoothing out any abrupt amplitude transitions from one speech frame to another.

In reality, though, forensic scientists do not need to understand the specifics of the FL process, but rather they need to have an overall appreciation of how much of the speech waveform, and in what respects, it might have been changed during transmission. The mechanism associated with detection of corrupted frames is quite complicated, but in simple terms, it looks for invalid speech parameter values as well as invalid transitions between frames. This process is normally referred to as sanity check in the CDMA network. Examples of the parameters checked are Linear Spectral Pairs (LSPs) and pitch delay (or pitch period). The LSP parameters are derived from a 10th order Linear Prediction Coding (LPC) analysis. LSPs are less sensitive to quantization noise which makes them superior to LPCs for transmission. The EVRC decoder applies a set of rules to determine whether the LSP parameters are valid or not. For example, if the received frame was originally coded at the highest bit rate (i.e., 8.55 kbps) and the maximum value of LSPs (1 and 2) is higher than the minimum of LSPs (4, 5, and 6), then this frame would be declared as lost or corrupted. This is because the LSP parameters must conform to an ascending order. Another aspect checked by the EVRC decoder is invalid transitions. If a frame is encoded with PPP at full rate (i.e., 8.55 kbps) and this was preceded either by a silence frame (0.8 kbps) or a 2 kbps frame, then this frame will also be treated as lost. Further details on sanity check can be found in 3GPP2 [5].
The lost-frame strategy in the EVRC is referred to as Frame Erasure. It involves replacing ‘Bad’ frames with ‘Good’ frames using previously received speech data. However, an artificially created ‘Good’ frame is not necessarily at the same bit rate as the ‘Bad’ frame it replaces. Indeed, as will be explained, it is usually at the highest bit rate of 8.55 kbps. To illustrate the various aspects of the lost-frame strategy of the EVRC [114], an example is discussed in this section. With this example, in order to convey the broad aspects of a process which in reality is very complicated, a distinction is made between data in a frame that could be classified as speech data (i.e., spectral shaping, voiced/unvoiced classification, pitch, etc.) as opposed to data related to amplitude. Figure 5.6(a) shows a sequence of received frames, four ‘Good’ (labelled with a superscript G in the figure) and three ‘Bad’ (labelled with a superscript B in the figure). The subscripts refer to the speech frame type such as silence (identified with S) or active speech (identified with an associated bit rate, namely 2, 4 or 8.55 kbps). Figure 5.6(b) shows the resulting speech frames that would be used to generate the decoded speech waveform. The superscript R associated with an individual frame identifies it as a replacement frame and a subscript has the same meaning as in Figure 5.6(a).

Frames 1 and 2 are both ‘Good’ and therefore remain unchanged. Frame 1 is silence and Frame 2 is active speech at one of the three bit rates, namely 2, 4 or 8.55 kbps. Frame 3 is ‘Bad’ and is replaced by an artificially generated frame at a bit rate of 8.55 kbps. Essentially the speech data used in the new Frame 3 is the same as that in the last ‘Good’ speech frame, namely Frame 2, except for a possible modification needed to correct for any change in bit rate between the two frames. If the bit rate of Frame 2 is 8.55 kbps, then the speech data in the new Frame 3 will be the same as that of Frame 2. If the bit rate for Frame 2 is either 2 or 4 kbps, a sophisticated bandwidth expansion of this speech data, specifically the LSP parameters, is performed to match the higher bit rate of the new Frame 3. As far as amplitude data for the new Frame 3 is concerned, this is made the same as for Frame 2. Thus the new Frame 3 in Figure 5.6(b) is identified as \(3^{R}_{8.55}(2)\). Frame 4 is also ‘Bad’ and its speech data would be replaced in an identical manner to Frame 3 (i.e., based on the speech data from the last ‘Good’ frame, namely Frame 2, but again with a possible bandwidth expansion). However, unlike for the new Frame 3, there would be an associated reduction in amplitude by a factor of 0.75 because Frame 4 is the second ‘Bad’ frame in
a sequence. Thus the new Frame 4 in Figure 5.6(b) is identified as \(4_{8.55}^R(2/)\). (Note: if a sequence of frames is ‘Bad’, the same process would be repeated, but with the amplitude of all subsequent replaced frames being reduced by a factor of \((0.75)^{(N-1)}\), where \(N\) is the consecutive ‘Bad’ frame number \((N \geq 2)\). Frame 5 is ‘Good’, so remains essentially unchanged, except for its associated pitch parameter. Again with the goal of minimizing discontinuities in the speech signal, in this case with respect to pitch, the pitch information of Frame 5 would be altered to become essentially an interpolation between the pitch of Frame 2 (and thus of the new Frames 3 and 4 which would have the same pitch as Frame 2) and that of Frame 5. Thus in Figure 5.6(b) the new Frame 5 is labelled as \(5_{8.55}^R(2,5)\) to indicate that it has derived its speech data from Frames 2 and 5). Frame 6, which is a silence frame, is also ‘Good’. However, one of the rules associated with the lost-frame process of the EVRC is that a silence frame cannot be preceded by a replaced frame that is high quality (i.e., a frame with a bit rate of 8.55 kbps). So Frame 6 is discarded and is replaced by a copy of the previous frame, namely the new Frame 5. It is thus labelled as \(6_{8.55}^R(2,5)\) in Figure 5.6(b). Finally, Frame 7 in this example is ‘Bad’ and so is replaced by essentially a copy of the ‘Good’ Frame 6 that was received, the only modification being with respect to its amplitude, this being recalculated slightly differently to other frames using procedures outlined in 3GPP2 [5], because it was preceded by a silence frame. The new Frame 7 then becomes \(7_{5}^R(6/)\) in Figure 5.6(b).

Another special case, that has not been shown in the figure, is when the current frame is an 8.55 kbps ‘Good’ frame following a 2 kbps PPP ‘Good’ frame which follows a ‘Bad’ frame. In this case, the pitch delay contour of the current frame is reconstructed by warping the last valid pitch delay (prior to the bad frame) and copying it to the previous 2 kbps PPP frame. This new pitch contour is then interpolated to produce the pitch delay contour of the current ‘Good’ frame coded at 8.55 kbps. The term pitch delay contour refers to a set of pitch periods and gains calculated for 5 ms segments of the 20 ms speech frame. Hence, four subframes (5 ms each) are processed when estimating the pitch delay contour.

It can be seen from the example of Figure 5.6 that though there were four ‘Good’ frames and three ‘Bad’ frames in the received seven-frame sequence of Figure 5.6(a), this has resulted in only two of these ‘Good’ frames, together with five artificially generated replacement
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.6: Illustration of the lost-frame mechanism implemented by the EVRC in the CDMA network. (a) A set of received speech frames, (b) Resulting set of speech frames used to reconstruct the speech signal.

frames, being used to generate the decoded speech waveform. It is also clear from this example that a considerable degree of sophistication has been incorporated into the lost-frame strategy of the EVRC, the underlying goal being to conceal as far as possible, from a perceptual standpoint, that data has been lost or corrupted during transmission. The unfortunate consequence from the standpoint of an FVC analysis is that determining from the recovered speech signal when this process has occurred is likely to be very challenging, if not impossible.

It is important to discuss now how the temporal location, together with the associated lost-frame corrective mechanism that it would have triggered, might impact upon the decoded speech waveform, both in terms of the time and spectral domains. To illustrate this aspect, a set of time waveforms and spectrograms have been produced for a token of the vowel /aI/ coded with the EVRC. A single lost frame has been introduced between the encoder and decoder, but at three different temporal locations, namely at Frames 3, 4 and 5.
The corresponding time waveforms and spectrograms of these are shown in Figures 5.7 and 5.8, respectively, with speech coded at anchor operating point OP0. The top sub-figure in Figure 5.7 shows 9 frames of the time waveform of the vowel segment (i.e., 180 ms of speech) without FL. The remaining sub-figures show the resulting time waveforms for Frames 3, 4 and 5 being lost, respectively. Figure 5.8 shows the spectrograms corresponding to the time waveforms shown in Figure 5.7. Clearly, the loss of a single frame can have quite an impact on the decoded speech frames that follow it. Moreover, this impact depends very much on exactly which frame is lost.
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.8: Spectrograms of the time waveforms shown in Figure 5.7 for the EVRC. (Dashed lines show the frame boundaries).

5.2.3 Background Noise (BN) at the transmitting end

BN at the transmitting end is frequently present in mobile phone communications. This situation becomes even more marked when there is a greater distance between the mobile phone microphone and the speaker, such as is likely to be the case with hands free terminals. The CDMA network incorporates a unique strategy, namely Noise Suppression (NS), which subtracts the BN present in every input speech frame [115, 116]. This process is primarily used to aid the more accurate classification of speech frames into voiced, unvoiced or transient prior to the coding stage. It should be noted that the NS process is preceded by a highpass filter which has a 3 dB cutoff frequency at about 120 Hz and a slope of about 80 dB/oct. This also assists in removing a large part of the low frequency BN without greatly affecting speech quality.

NS is repeated twice for every frame (i.e., every 10 ms) and uses a set of energy estimators and voice metrics to determine characteristics of the noise signal, and thus assist in its subsequent removal [114, 117]. This 10 ms section of the input speech frame is first
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

windowed using a smoothed trapezoidal window and then transformed into the frequency domain using a 128-point Fast Fourier Transform (FFT). The resultant 128 frequency bins are then grouped into 16 bands (or channels), which approximate the ear’s critical bands. The energy present in each critical band is estimated by averaging the magnitude of all frequency bins within this band [26] (see Equation 5.2).

\[
E_{ch}(m,i) = \frac{1}{f_H - f_L + 1} \sum_{k=f_L(i)}^{f_H(i)} |G_m(k)|
\]  

(5.2)

where \(E_{ch}\) is the channel energy, \(m\) is index of a speech frame, \(i\) is the channel number \((i = 0, 1, 2 \ldots 15)\) and \(f_H\) and \(f_L\) are the highest and lowest frequency bins of channel \(i\), respectively. \(|G_m(k)|\) is the magnitude of the \(kth\) frequency bin in that particular channel. The channel noise energy \(E_N\) is estimated in a similar way to \(E_{ch}\), but from pauses that naturally occur in human speech. This is then combined with the channel energy to determine the Signal-to-Noise Ratio (SNR) of each channel \((SNR_{ch})\) (see Equation 5.3). The SNR is primarily used as a voice metric to determine if the current frame contains active speech or only noise. If this frame is considered noise, the current \(E_{ch}\) will be used to update the channel noise estimator \(E_N\) according to the following formula (5.4):

\[
SNR_{ch}(i) = 10 \log_{10}(\frac{E_{ch}(m,i)}{E_N(m,i)}),
\]  

(5.3)

\[
E_N(m + 1, i) = \alpha E_N(m, i) + (1 - \alpha) E_{ch}(m, i),
\]  

(5.4)

where \(m + 1\) indicates that this new \(E_N\) estimator will be used for the next speech frame, \(\alpha = 0.9\) is the channel noise smoothing factor.

Despite the beneficial effects of NS in removing BN, it can add distortion to the coded speech signal when BN levels are high. The most common types of BN in mobile phone communications are babble\(^3\), car and street noises, and these noise types have therefore been used in these experiments. Typical SNR levels at the transmitting end vary from 9 to 20 dB [118]. It needs to be mentioned here that the impact of NS on the speech signal

---

\(^3\)confused sound or background speech of a group of people talking simultaneously.
varies depending on the anchor operating point being used. For instance, OP0 uses only 8.55 kbps CELP for coding speech vowels, whereas the other OPs use a combination of CELP and PPP. As previously discussed in Section 5.2.1, with OP0, a set of three voiced frames is encoded at the same highest bit rate (8.55 kbps), whereas this same set would be encoded at a deterministic pattern of bit rates in the cases of OP1 and OP2. The use of the PPP algorithm in OP1 and OP2 can mask the effect of NS because this coding technique repeats information from previous frames if no significant changes are found between the current and previous frames.

In order to examine how the coding algorithm, together with NS, might impact upon the speech signal, a set of time waveforms and their respective spectrograms were produced for a token of the word ‘eight’ as shown in Figures 5.9 and 5.10, respectively. This token, however, is different to the one used in Section 5.2.1, when examining impact of DRC on speech, but it belongs to the same speaker. The rationale for not using the same token in this experiment is that the impact of NS became much clearer when a another token was used.

Figure 5.9(a) shows the time waveform of the original speech segment. Figure 5.9(b) shows the same speech segment with street noise being added at SNR = 6 dB. The noisy speech has then been coded at OP0 and OP2 (see Figures 5.9(c) and 5.9(d), respectively). The coding pattern in the EVRC for a sequence of three voiced frames is PPP, PPP and CELP for OP2 (Figure 5.9(d)), whereas it is always CELP for OP0 (5.9(c)). See Sections 5.1.4 and 5.2.1 for further details about the coding patterns used in the EVRC.

Examination of Figures 5.9 and 5.10 show that the EVRC has done a good job in reducing BN despite the high levels of noise being added to the speech signal. However, speech signal characteristics have changed slightly as a result of the NS process. In the case of OP0 (Figure 5.9(c)) a large part of the original speech in Frame 3, together with the BN originally present there, have been removed. Interestingly, this was not the case for OP2, where Frame 3 remained intact after NS, though, OP2 is expected to produce lower speech quality than OP0. The reason for this is again the use of PPP in OP1 and OP2, which exploits the fact that pitch contours do not significantly change over two or more consecutive frames. Hence, it repeats information from previous speech frames and this can mask the damage.
caused by NS. Therefore, one might expect a poorer performance of an FVC when using high-quality coded speech in the presence of high BN levels (i.e., low SNRs). Given that in the forensics arena it is quite common to be working with speech evidence contaminated by high levels of BN, the fact that the NS might be making this situation even worse is clearly of concern.

Figure 5.9: A set of time waveforms showing the impact of BN on the word “eight”. (a) Clean and uncoded speech, (b) uncoded speech with added street noise at SNR = 6 dB, (c) EVRC response to BN using OP0, (d) EVRC response to BN using OP2. (the red dashed lines show frame boundaries of the vowel /aI/ in eight).
5.2 Ways in which the CDMA mobile phone network impacts on the speech signal

Figure 5.10: A set of spectrograms showing the impact of BN on the word “eight”. (a) Clean and uncoded speech, (b) uncoded speech with added Street noise at SNR = 6 dB, (c) EVRC response to BN using OP0, (d) EVRC response to BN using OP2.

5.2.4 Silence Frames (SFs)

With mobile phone networks, reducing the bit rate as much as possible is of great importance. In addition to the processes described previously with respect to dynamic rate coding, bit rate is reduced further by the manner in which SFs are handled. Though referred to as SFs, they are in fact frames containing BN. Studies have shown that from the standpoint of the comfort of users, transmission of frames containing BN is important. At the transmission end, SFs (i.e., frames not containing speech) are detected using a Voice Activity Detection (VAD) mechanisms. A frame of BN is then transmitted to the receiving mobile. When subsequent SFs are detected, they are not transmitted to the receiver [5]. Instead, the codec at the receiving end is notified of the SF and it then inserts the representative SF that it received previously. The representative SF is referred to as comfort noise (CN). Provided that the BN characteristics at the transmission end do not change, the representative SF at the receiving end continues to be used. Only when the BN characteristics
at the transmission end change significantly will this representative SF at the receiving end be updated. A further complication arises when the VAD detection makes errors [4]. In other words, it wrongly classifies a frame of speech as silence. If this happens, segments of the speech signal will be replaced by previously sent samples of BN. From the standpoint of forensic speech analysis, it is therefore important to realize that BN segments are not transmitted to the receiving end with any degree of fidelity. So, any attempt to extract useful forensic information from BN segments should be undertaken with understanding of the above processes. The impact of the CDMA network on SFs has not been investigated in this study, as focus has been on the active speech frames.

5.2.5 All aspects combined

Figure 5.11 illustrates how the CDMA network might impact on the speech signal in regard to the four previous aspects combined. Figure 5.11(a) shows the original speech signal at the sending end, with Figure 5.11(b) showing the recovered speech signal at the receiving end. In Figure 5.11(a), the speech signal has been segmented into frames of 20 ms and the first frame has been correctly classified as silence. Frames 2 to 8 have been detected as speech and coded at different bit rates. In this case, the 2nd frame has been correctly classified as speech. The presence of NS in the EVRC helped to improve this classification. Frame 9 has been correctly detected as a SF. At the receiving end (Figure 5.11(b)), Frames 1 and 9 have been replaced with the representative comfort noise frame. Frames 2 to 5 were received correctly and decoded. However, in this example, Frames 6 and 7 were lost and the FL mechanism has been used to replace them with a modified version of Frame 5. Since the bit rate of Frame 5 is 4 kbps, a bandwidth expanded version of the speech data in this frame has been used for the new Frame 6 in order to match the higher bit rate of replaced frames (i.e., 8.55 kbps). This frame is identified as $6_{8,55}^R(5)$. The amplitude data for Frame 6 is made to match Frame 5. Frame 7 is also identified as ‘Bad’ and its speech data has been replaced in the same manner as Frame 6. However, its amplitude is reduced by a factor of 0.75 because Frame 7 is the second ‘Bad’ frame in a sequence and the resultant frame is identified as $7_{8,55}^R(5/5)$. Frame 8 has been successfully received and decoded. However, a
large part of the original speech present in this frame has been removed due to the noise subtraction process inherent with the EVRC.

5.3 Chapter summary

This chapter presented key features of the CDMA mobile phone network with respect to its potential impact on the speech signal and consequently the outcome of an FVC analysis. Two different approaches can be used for assessing the impact of mobile phone networks on speech. One approach is to transmit speech through an actual network, but this can only encompass a small subset of the actual scenarios in the network. A better strategy is to pass speech files through the speech codec and drive it in a manner similar to an actual mobile phone network. However, a thorough understanding of the rules in the network under which the codec modes might be initiated is essential to this approach.

There are three major aspects of the CDMA network, which can negatively impact the quality of “active” speech: DRC, FL and BN at the transmitting end. DRC is the process of instructing the EVRC to switch between different anchor operating points in response to changes in capacity conditions. These conditions can be broadly classified into three categories, namely high, medium and low, which in turn translate into low, medium and high-quality coded speech, respectively. The second aspect is FL, where this mechanism in the codec has been carefully designed to minimize the speech artefacts arising from lost or corrupted frames. This means identifying lost frames in a mobile speech sample is going to be quite challenging, if not impossible. The temporal location of lost frames can also have quite a significant impact on the recovered speech signal, specifically when losses occur around the transition region of vowels. BN picked up by the mobile microphone can also directly impact the coding process of speech signals. Though the EVRC incorporates a mechanism for NS, this task becomes quite challenging to the codec when BN levels are high. Although the AOPs OP1 and OP2 produce lower-quality speech than OP0, their use of the PPP algorithm can mask the damage caused by NS and thus might improve the FVC results. Lower-quality speech is normally the case when calls are made from a highly populated area such as a city.
5.3 Chapter summary

Figure 5.11: Impact of the CDMA network on speech. (a) processes implemented in sender mobile; (b) resulting processes in receiver mobile. In this figure it is assumed that Frames 6 and 7 are lost and that at the receiving end they are replaced by repeats of Frame 5. With respect to Frame 8 at the receiving end, the NS mechanism in built into the EVRC has removed part of the original speech signal.
It needs to be noted that the resultant speech segments in Section 5.2 have been inspected by listening. The impact of BN is relatively easier to detect by ear, as compared to DRC and FL, especially at SNRs less than 9 dB. With respect to DRC and FL, it was hard to detect the subtle differences caused by these aspects by mere listening, especially when using short segments of speech. Further, the codec has been designed with quality in mind and as much as possible it reduces discontinuities and speech artefacts, which makes it even harder to detect such changes and impacts.
Chapter 6

The Comparative Performance of Automatic Acoustic Parameters for FVC

In this chapter the performance of various acoustic feature sets is investigated to determine the best performing feature set when using CDMA mobile phone speech [108]. The experimental methodology together with a description of the speech database used in the these experiments is provided first. Subsequent sections show the FVC performance results of these feature sets using first uncoded speech and then EVRC-coded speech.

6.1 Experimental methodology

The XM2VTS database [26] was used for all the experiments in this chapter\(^1\). This is a multimodal database containing digital speech recordings of 295 speakers along with synchronized video recordings. The language of the recordings is English with a Southern British accent. The speakers were recorded on four different occasions at intervals of one month. During each session each speaker repeated three sentences twice. The first two

---

\(^1\)This database has also been used for the investigation in all subsequent chapters.
sentences included reading a random digit sequence and the last recording is a phonetically balanced sentence. The sentences are as follows: 1. “zero one two three four five six seven eight nine”, 2. “five zero six nine two eight one three seven four”, 3. “Joe took father’s green shoe bench out”. The speech files are sampled at 32 kHz with 16 bit digitization. Of the 295 speakers, 156 are male speakers and only those have been considered in these experiments. Of the male speakers, 26 candidates were removed from the outset as their recordings were judged to be less audible or they appeared to have accents quite different to the others.

Three words; “nine”, “eight” and “three”, were extracted from the speech of the 130 speakers using the audio editing software Goldwave [119] and Wavesurfer [120]. After extracting the words of interest, the vowels /aI/, /eI/ and /i/ were extracted from their respective words by following a careful auditory and acoustic procedure\footnote{The respective spectrograms and time waveforms of the extracted vowels were examined and then inspected by listening.} [13]. Even though the database contains four different recording sessions, only three of them have been used in these experiments. In summary, three non-contemporaneous sessions have been used with three vowels per session and four tokens per vowel. The selected vowels include two diphthongs and a monophthong.

The speech database of 130 speakers was divided into three groups: 44 speakers in the background set, 43 speakers in the development set and 43 speakers in the testing set. With 43 speakers in the testing set, 43 same-speaker comparisons and 903 different-speaker comparisons are possible. Two same-speaker comparison results were obtained for each speaker in the testing set by comparing their first recording session with their own recordings in sessions 2 and 3. Three different-speaker comparisons were produced for each speaker by comparing their first recording session with other speakers’ recordings from sessions 1, 2 and 3. The background set remained the same for all comparisons and contained two recording sessions for all 44 speakers. Note that the sole purpose of the development set is to train the logistic regression-fusion system, the resulting weights of which are then used to combine LRs calculated from individual vowels for each comparison in the testing set.
6.1 Experimental methodology

Figure 6.1 shows a block diagram of the experimental procedure followed in this chapter. For purposes of comparison, two separate sets of experiments have been considered: the first using speech that has not been processed by the codec, and the second using speech that has, and this under two operational modes of the EVRC. These experiments involve a comparison, in terms of $C_{llr}$ and $CI$, between five types of speech parameters (details of which are found in Appendix A).

A set of CCCs, RCCs, LPCCs and MFCCs extracted from the entire vowel segment have been considered. With FTs, a vowel segment has been segmented into 20 ms Hamming windowed frames. With respect to the number of parameters used for each feature set, it was considered to be important in this investigation to use as many parameters as possible in a feature set in order to ensure fairness when comparing results between feature sets. For the first experiment, a total of 100 CCCs were considered (50 causal and 50 anti-causal components). This number was chosen based upon previous research examining the use of CCCs when applied to speech [40]. This is also motivated by the fact that the overall vocal tract impulse response (i.e., $\hat{h}(n)$ - see Appendix A) are dominant in this region. For the second experiment (with RCCs), only the first 50 coefficients were used. This number was chosen because RCCs are the even component of CCCs and contain no anti-causal components. 16 LPCCs were used in the third experiment (a tenth order LPC analysis has been used). The choice of 16 LPCCs was motivated by experiments in this study with mobile phone speech [108], which have shown that this number of coefficients, when extracted from a 10th order LPC analysis, generally give optimum FVC performance. For the fourth experiment, 23 MFCCs were used (all DCT coefficients have been considered here). This is the maximum number of MFCCs that can be extracted, this being determined by the data’s sampling frequency [40]. In these experiments, the speech data was downsampled to 8 kHz, which is the standard value used in the EVRC. At 8 kHz sampling frequency, a maximum of around 23 MFCCs can be extracted. For the last experiment, a fourth order DCT function was applied to each trajectory of the first three formants (i.e., a total of 12 parameters). Refer to Appendix A for further details.

The EVRC-coded speech experiment was repeated twice, once using OP0 (high-quality speech) and the other using OP2 (low-quality speech). The choice of different modes in the
EVRC was motivated by the fact that each anchor operating point incorporates different coding techniques [118], such as CELP, NELP and PPP, and these may well have different impacts on the speech parameters of interest. As discussed in Chapter 5, the CDMA network changes the mode of operation in response to increases (or decreases) in the number of users accessing the network (an aspect being thoroughly investigated in Chapter 7), but it was decided to use fixed modes of operation (i.e., OP0 and OP2) for the experiments in this chapter. The rationale for this was that one should keep other factors constant when investigating a certain aspect of an experiment. The background set used in the CDMA experiments contained speech coded using the specific mode being investigated. In other words, there was no mismatch between the background set and the speakers involved in the comparison.

Figure 6.1: Block diagram showing the experimental procedure used to investigate the performance of FVC using mobile phone speech

The FVC results in all experiments were computed using feature sets extracted from tokens of the vowels: /aI/, /eI/ and /i/. The speech parameters of each feature set were then used
as input to PCAKLR and their resulting transformed parameters being used in the FVC analysis. The results from the individual vowels were then fused using logistic regression, the associated weights being determined from the 43 speakers in the development set. With three recording sessions, two LRs were calculated for every same-speaker comparison and three LRs for every different-speaker comparison. The mean LR value was then computed for each comparison (i.e., from two same-speaker LRs and three different–speaker LRs). The accuracy of the resulting mean LRs was computed using $C_{llr}$. The $CI$ was estimated for each comparison and the reliability has then been expressed in terms of the average of these $CI$ values. The final results are plotted using Tippett plots.$^3$

## 6.2 Performance of uncoded speech

Table 6.1 compares the resulting FVC performance for the uncoded experiments described. On the basis of both $C_{llr}$ and $CI$, it is clear that CCCs (i.e., Experiment 1) outperformed other feature sets. A graphical presentation of the coded speech results is shown in Figure 6.9. In terms of accuracy, RCCs come next followed by MFCCs, LPCCs and then FTs. The last three feature sets tend to have larger variation (i.e., worse precision) than CCCs and RCCs. This is evident by the higher $CI$ values of these sets.

Table 6.1: Performance of various speech parameters using uncoded speech

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Parameter</th>
<th>$C_{llr}$</th>
<th>$CI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CCCs</td>
<td><strong>0.099</strong></td>
<td><strong>1.523</strong></td>
</tr>
<tr>
<td>2</td>
<td>RCCs</td>
<td>0.139</td>
<td>1.707</td>
</tr>
<tr>
<td>3</td>
<td>MFCCs</td>
<td>0.167</td>
<td>2.299</td>
</tr>
<tr>
<td>4</td>
<td>LPCCs</td>
<td>0.218</td>
<td>2.214</td>
</tr>
<tr>
<td>5</td>
<td>FTs</td>
<td>0.291</td>
<td>2.132</td>
</tr>
</tbody>
</table>

The results have been further analyzed using Tippett plots. Figure 6.3 shows a Tippett plot for the best performing feature set (the CCC results). Figures 6.4, 6.5, 6.6 and 6.7 shows the corresponding Tippett plots for RCCs, MFCCs, LPCCs and FTs, respectively.$^3$

---

$^3$This same procedure was also used for all experiments in Chapters 7 to 10.
6.2 Performance of uncoded speech

Figure 6.2: $C_{llr}$ v.s. $CI$ for uncoded speech for various speech parameters.

In the Tippett plots shown, the solid blue curve rising towards the right represents the same-speaker comparison results and the solid red curve towards the left represents the different-speaker comparison results. The dashed lines on either side of the solid curves represents the variation found in a particular LLR (i.e., $CI^4$).

It would appear from Figures 6.5 to 6.6 that MFCCs and LPCCs have outperformed CCCs, as the same- and different-speaker curves are farther compared to CCCs. However, the performance in the vicinity of the LLR= 0 decision boundary is more important than the performance of higher magnitude LLRs. As can be seen from these figures, CCCs have outperformed all other feature sets in this region for the same-speaker comparisons. However, RCCs and MFCCs have shown similar performance to CCCs in terms of different-speaker comparisons. The proportion of same-speaker misclassifications was lower for CCCs (almost none) compared to RCCs and MFCCs.

---

$^4$The $CI$ results shown in Table 6.1 are calculated by taking the mean value of all individual $CIs$ for both the same- and different-speaker comparisons.
6.2 Performance of uncoded speech

Figure 6.3: Tippett plot showing FVC performance of CCCs using uncoded speech.

Figure 6.4: Tippett plot showing FVC performance of RCCs using uncoded speech.
6.2 Performance of uncoded speech

Figure 6.5: Tippett plot showing FVC performance of MFCCs using uncoded speech.

Figure 6.6: Tippett plot showing FVC performance of LPCCs using uncoded speech.
Figure 6.7: Tippett plot showing FVC performance of FTs using uncoded speech.

The APE-plot in Figure 6.8 shows that the improvement in $C_{llr}$ when using CCCs is attributable to the improved discrimination performance of this feature set (the green portion $C_{llr_{min}}$).
Figure 6.8: APE plots showing the losses in $C_{llr}$ using various speech parameters extracted from uncoded speech.

On the other hand, the calibration performance (the red portions $C_{llr_{cal}}$) for all these parameters is comparable except for FTs which have exhibited higher losses in terms of $C_{llr_{cal}}$.

### 6.3 Performance of EVRC-coded speech

Table 6.2 represents the impact of the EVRC on FVC performance using the previous feature sets in terms of $C_{llr}$ and $CI$. The same set of experiments was repeated for the two anchor operating points, OP2 (corresponding to low-quality coding) and OP0 (corresponding to high-quality coding). The results for uncoded speech are also included in this table for the ease of comparison. A graphical presentation of the coded speech results is shown in Figure 6.9.
6.3 Performance of EVRC-coded speech

Table 6.2: Performance of various speech parameters using EVRC-coded speech

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Parameter</th>
<th>OP0</th>
<th></th>
<th></th>
<th>OP2</th>
<th></th>
<th></th>
<th>Uncoded</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CCCs</td>
<td>0.135 &lt;br&gt; 2.265</td>
<td>CI</td>
<td>0.106 &lt;br&gt; 2.048</td>
<td>CI</td>
<td>0.099 &lt;br&gt; 1.523</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>RCCs</td>
<td>0.151 &lt;br&gt; 1.825</td>
<td>CI</td>
<td>0.135 &lt;br&gt; 1.976</td>
<td>CI</td>
<td>0.139 &lt;br&gt; 1.707</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MFCCs</td>
<td>0.122 &lt;br&gt; 1.861</td>
<td>CI</td>
<td>0.127 &lt;br&gt; 2.036</td>
<td>CI</td>
<td>0.167 &lt;br&gt; 2.299</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>LPCCs</td>
<td>0.249 &lt;br&gt; 1.644</td>
<td>CI</td>
<td>0.285 &lt;br&gt; 1.674</td>
<td>CI</td>
<td>0.218 &lt;br&gt; 2.214</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>FTs</td>
<td>0.401 &lt;br&gt; 1.304</td>
<td>CI</td>
<td>0.373 &lt;br&gt; 1.919</td>
<td>CI</td>
<td>0.291 &lt;br&gt; 2.132</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For all feature sets, a comparison between the EVRC-coded and uncoded speech results show that the EVRC negatively impacts upon FVC accuracy (i.e., $C_{llr}$), except for MFCCs, which have actually improved. With respect to the coded speech results, the accuracy improves as the speech quality improves (OP0 in this case) for MFCCs and LPCCs, but it is marginally worse for CCCs, RCCs, and FTs with high-quality speech. This behaviour is unexpected, as one expects performance to improve when using higher-quality of speech coding (it is not clear at this stage why this might be so).

There are a number of reasons for the degradation in accuracy using CDMA mobile phone speech. The EVRC is designed to achieve the best speech quality at relatively low bit rates. To do this, the speech coding process ignores information about glottal shaping and lip radiation filters (details of the source–filter model are discussed in Appendix A) as these components only add a little improvement to the perceptual quality of speech. As a result, any speech parameter that uses information about these components will lose some of its discriminative power upon speech coding, examples of which are the CCCs and RCCs. This perhaps also explains the better performance of MFCCs, which are not directly related to any particular speech component, but rather roughly estimate the overall energy in each filter bank region.

In terms of reliability of the EVRC-coded results, the $CI$ has improved for MFCCs, LPCCs and formant trajectory (whereas the opposite has happened to CCCs and RCCs). This was the case for both speech coding qualities (OP0 and OP2). This is counter intuitive, as one might expect $CI$ to be worse for coded speech. However, the improvement in $CI$ is likely a result of the quantization processes inherent in the compression implemented by the EVRC. When speech parameters are quantized, small variations in parameter values are
removed, which in turn makes speech samples more similar, thus reducing, for example, within-speaker variation across different recording sessions. The quantization processes combined with using the logarithm operator in MFCCs extraction could be another reason for the improvement in CI. The logarithm is used to suppress small differences in energy (i.e., the logarithm grows slowly as the energy value increases). This process aligns with the perceptual aspect of the human ear. If a speech sound is quite loud to begin with, variations in energy at that level (and higher) would not sound very different.

The accuracy of FVC results was further investigated using Tippett plots. Figures 6.10 to 6.14 show the respective FVC performance of CCCs, RCCs, MFCCs, LPCCs and FTs when using high-quality coded speech (i.e., OP0).\(^5\) The Tippett plots of CCCs, RCCs and MFCCs have shown comparable performance in terms of both the same- and different-speaker results in the region around LLR = 0. However, the proportion of same-speaker

\(^5\)Similar behaviour was observed in the vicinity of LLR= 0 for the low-quality speech coding (i.e., OP2) and thus its corresponding Tippett plots have not been shown here.
misclassifications in the MFCCs experiment of coded speech (Figure 6.12) was significantly lower (almost none) than the uncoded speech experiment shown in Figure 6.5. With LPCCs, speech coding has marginally negatively impacted both same- and different-speaker comparisons (compare Figure 6.6 with Figure 6.13). With respect to the formant trajectory experiment, the accuracy is significantly worse compared to uncoded speech. This is evident in the higher proportion of different-speaker misclassifications seen in Figure 6.14 as compared to Figure 6.7. Surprisingly, the same-speaker comparisons have improved here. This behaviour is again likely to be a result of the quantization process associated with the coding of these parameters which reduces both the within- and between-speaker variations, leading to an improvement in the same-speaker comparisons, but worsening different-speaker results.

Figure 6.10: Tippett plot showing FVC performance of CCCs using high-quality speech coding (OP0).
6.3 Performance of EVRC-coded speech

Figure 6.11: Tippett plot showing FVC performance of RCCs using high-quality speech coding (OP0).

Figure 6.12: Tippett plot showing FVC performance of MFCCs using high-quality speech coding (OP0).
6.3 Performance of EVRC-coded speech

Figure 6.13: Tippett plot showing FVC performance of LPCCs using high-quality speech coding (OP0).

Figure 6.14: Tippett plot showing FVC performance of Formants using high-quality speech coding (OP0).
To further understand the $C_{llr}$ results, APE-plots have been produced for the high and low-quality speech coding experiments (Figures 6.15 and 6.16, respectively). As with the clean speech experiment (Figure 6.8), the $C_{llr}$ of coded speech is again dominated by the discrimination loss ($C_{llr_{\text{min}}}$). The discrimination loss marginally decreases with higher speech coding quality in the case of MFCCs and LPCCs. The opposite is observed for other feature sets, namely the $C_{llr_{\text{min}}}$ is higher when using high-quality coded speech. The improvement in $C_{llr}$ when using MFCCs is attributable to the improved discrimination performance of this feature set. Differences in the calibration loss ($C_{llr_{\text{cal}}}$) between the coded and uncoded speech were marginal.

Figure 6.15: APE plots showing losses in $C_{llr}$ using various speech parameters extracted from speech coded at high quality (OP0).
From the above results, MFCCs appear to be the best performing feature set when dealing with EVRC-coded speech and, therefore, it has been used in subsequent experiments to investigate the impact of different aspects of the CDMA network on FVC. CCCs and RCCs did not perform as well as MFCCs and this is probably due to the fact that the speech coding in the CDMA mobile phone network removes relevant information about the glottal shaping and lip radiation components of the speech signal. These components are essential to the performance of these parameters and removing them could be expected to result in reduced FVC performance. However, CCCs were still the second best performing feature set, followed by RCCs and then LPCCs. The performance of FTs was the worst of all parameters.
6.4 Chapter summary

A number of experiments have been conducted to assess the performance of various types of cepstral coefficients, as well as FTs, when CDMA mobile phone speech is used as opposed to clean, uncoded speech. MFCCs were found to be the best performing feature set in terms of the accuracy and precision of LR results. Though CCCs have slightly outperformed MFCCs in one instance, MFCCs have the advantage of being more robust to speech quality changes than CCCs. This is evident in the lower $CI$ value of MFCCs. RCCs did not perform as well as MFCCs and this is likely linked to the fact that the encoding process removes relevant information about the glottal shaping and lip radiation components of speech. As these components are essential to the performance of RCCs and CCCs, removing them could be expected to result in a worse FVC performance. LPCCs and FTs were the worst performing of all the feature sets.
Chapter 7

Impact of Dynamic Rate Coding (DRC) on FVC Analysis

There are three key aspects of the CDMA network which ultimately impact upon speech quality, namely (i) Dynamic Rate Coding (DRC), (ii) Frame Loss (FL) and (iii) Background Noise (BN). This chapter focuses on the first of these aspects (i.e., DRC) in isolation to others. The subsequent two chapters describe the experimental procedure used to investigate the other aspects and then discuss their respective results.

7.1 Experimental methodology

The experimental procedure in this section involves a comparison in terms of $C_{llr}$ and $CI$ between two FVC analyses, the first using uncoded speech, the second using speech that has been coded by the EVRC, and this under representative operational modes of the codec. As discussed earlier in Section 5.2.1, adjusting the average data rate (ADR) in the CDMA network is done in accordance with the number of users accessing the system, referred to here as channel capacity. Upon changes in this capacity, the network instructs the EVRC to meet some target ADR, this being in the range 4.8 kbps to 9.6 kbps. With a continuous range of ADRs, there is a large number of possibilities with correspondingly different impacts upon speech quality, and thus FVC. Rather than conducting an FVC experiment at
every one of these, the focus has been on three channel capacity scenarios, high, medium and low, which in turn will result in low-, medium- and high-quality speech, respectively, as shown in Figure 7.1.

The CDMA-codec platform\(^1\) developed in this study has been used to create ADR files in accordance with the three speech qualities (or capacity conditions) discussed above. The ADR values are automatically generated upon specifying the speech quality desired. For each speech quality, the ADR values (see Figure 5.2) are sampled at different points of the continuous range according to the following: (i) lower-half (low-quality speech), (ii) entire range (medium-quality speech) and (iii) upper-half (high-quality speech). The ADR values chosen for each of the ranges are shown in Figure 7.1.

Figure 7.1: Block diagram illustrating how speech files have been processed corresponding to different capacity conditions.

In the high capacity scenario, the ADR values are selected from anchor operating point regions OP1 and OP2 corresponding to bit rates from the set 4.8, 5.8, 6.2, 6.6 kbps. For medium capacity, they are selected from OP0, OP1 and OP2, corresponding to bit rates

\(^{1}\)The design of this platform is based on an understanding of the aforementioned aspects and how they are linked to what is happening in the network as a whole (see Appendix B).
7.1 Experimental methodology

from the set 4.8, 5.8, 6.2, 6.6, 7.0, 7.5, 8.5, 9.6 kbps. In the low capacity scenario, they are selected from anchor operating point regions OP0 and OP1 corresponding to bit rates from the set 7.0, 7.5, 8.5, 9.6 kbps. In all three cases the actual ADR value used for a particular speech frame in these experiments was randomly selected from the groupings listed above according to a uniform distribution. In the absence of any information pointing to some other distributions, this was considered the most reasonable choice.

For example, consider a speech token that has a duration of eight frames (160 ms) and the capacity condition is chosen to be high. The ADR values will firstly be selected from the set 4.8, 5.8, 6.2 and 6.6 kbps. Next, an ADR entry will be generated for each speech frame, these being randomly selected from the above set according to a uniform distribution. The EVRC will then be instructed to meet the ADR values specified, which is likely to result in switching between OP1 and OP2. Upon selecting an anchor operating point, the codec will use an internal selection mechanism to decide on an actual bit rate for each speech frame. In summary, the bit rate chosen for a particular speech frame is a function of the speech characteristics present in that frame as well as the anchor operating point chosen (see Figure 5.2).

The selection of discrete ADR values in these experiments ensures better coverage of the three channel capacity ranges and thus enforces switching between anchor points in order to reflect more realistic scenarios in the network. For instance ADR values for medium speech quality (medium channel capacity in this case) can take on values between 4.8 and 9.6 kbps [5], but random selection from such a continuous range can result in these being more concentrated at one part of the range than another. Using discrete values increases the likelihood of switching between anchor points and ensures a better coverage of the range.

The FVC experimental procedure discussed in Section 6.1 has been used for these experiments. As mentioned earlier, MFCCs were considered for all experiments investigating impact of the CDMA network on FVC. Further, the speech quality used for speakers involved in the comparison was the same as that of the background set (i.e., matched conditions were considered).
7.2 FVC performance for low-, medium- and high-quality coded speech

Table 7.1 shows the FVC performance of the different coded speech qualities incorporating DRC in terms of $C_{llr}$ and $CI$ and these are compared with that of uncoded speech. The accuracy and reliability of the FVC results for the uncoded speech experiment using MFCCs have been already presented in Section 6.2.

<table>
<thead>
<tr>
<th>Speech condition</th>
<th>$C_{llr}$</th>
<th>$CI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncoded</td>
<td>0.167</td>
<td>2.299</td>
</tr>
<tr>
<td>Low quality</td>
<td>0.152</td>
<td>1.702</td>
</tr>
<tr>
<td>Medium quality</td>
<td>0.135</td>
<td>1.803</td>
</tr>
<tr>
<td>High quality</td>
<td>0.111</td>
<td>1.869</td>
</tr>
</tbody>
</table>

Figure 7.2 shows the results of Table 7.1 plotted graphically. As was the case with the fixed mode experiments (OP0 and OP2) previously discussed in Section 6.3, EVRC-coded speech has resulted in a better performance with respect to $CI$ than uncoded speech. Again, it is likely that parameter quantisation is the underlying cause of this. Differences in $CI$ values between the different speech qualities are marginal.

In terms of the accuracy of LR results, this was also better for coded speech than uncoded speech for two reasons. First, the quantisation process limits the within-speaker variation and thus makes their samples look more similar. This also explains the improved performance of the same-speaker comparisons, which is evident in the lower misclassifications of these comparisons as observed in the Tippett plots of Figures 7.3, 7.4 and 7.5. Second, MFCCs do not carry information about specific components of the source-filter model and, therefore, speech coding is unlikely to affect their discriminative power. In regard to the results of different speech qualities, accuracy was found to increase with increasing speech quality, which is obviously a trend one would expect.

Reminder: the lower the values of $C_{llr}$ and $CI$, the more accurate and precise, respectively, are the LR results.
The performance of coded speech results was further investigated using Tippett plots as shown in Figures 7.3, 7.4 and 7.5 for low, medium and high speech coding qualities, respectively. A comparison of Figure 7.4 with Figure 6.5 (the Tippett plot for uncoded speech) confirms that the improved accuracy of coded speech is a result of improved same-speaker comparisons.

The APE–plots of Figure 7.6 examine in more detail the $C_{llr}$ results of Table 7.1. These show that the discrimination loss ($C_{llr_{\text{min}}}$) is worst for uncoded speech and for coded speech it improves with increasing speech quality. The former observation is not at all expected, but the latter certainly is.

Another important aspect of the APE plots is the calibration loss ($C_{llr_{\text{cal}}}$) component of $C_{llr}$. The $C_{llr_{\text{cal}}}$ for high, medium and low qualities of speech are: 0.048, 0.056 and 0.048, respectively. These results are in fact comparable to the uncoded speech with $C_{llr_{\text{min}}} = 0.050$ which shows that the improvement in $C_{llr}$ for coded speech was a result of improved discrimination. Overall, though, it can be concluded from these experiments that the FVC performance is marginally better in terms of both the $C_{llr}$ and CI for EVRC-coded speech at different speech qualities as compared to uncoded speech using MFCCs.
7.2 FVC performance for low-, medium- and high-quality coded speech

Figure 7.3: Tippett plot showing FVC results for low-quality coded speech.

Figure 7.4: Tippett plot showing FVC results for medium-quality coded speech.
7.2 FVC performance for low-, medium- and high-quality coded speech

Figure 7.5: Tippett plot showing FVC results for high-quality coded speech.

Figure 7.6: APE-plot for various speech quality conditions.
Three experiments were conducted to examine the effects of DRC on the performance of FVC analysis. The DRC process is a function of the number of users accessing the network and thus it can be broadly classified into three capacity groups which translate into low-, medium- and high-quality speech. Discrete ADR values were selected for each speech quality and this step was necessary to ensure a better coverage for the ADR continuous range (from 4.8 to 9.6 kbps) and thus enforce switching between the AOPs in accordance with the chosen scenario.

Surprisingly, in virtually all cases, both the validity and reliability of the FVC results were better for coded speech than uncoded. The results were best for higher-quality speech, which was expected. The improvements in $CI$ and $Cllr$ are likely due to the quantisation process inherent in the speech codec, which reduces the within-speaker variation.

The next question is as to whether these three investigated scenarios could happen in reality. In fact, though the DRC was investigated in isolation from other aspects, it is still possible to have a call being made under such conditions from a place with low levels of BN present as well as high-SNR transmission, making the loss of frames unlikely. This could possibly be the case if a call was made from a quiet place (e.g., a house) that is located near to a base station.
Chapter 8

Impact of Frame Loss (FL) on FVC Analysis

The goal of this chapter is to study the impact of frame loss on FVC in isolation from BN and limiting the change in DRC as much as possible. As mentioned earlier in Section 5.1.3, DRC cannot be completely disabled in EVRC, but, rather, its operation can be restricted to one of the AOPs. This ensures that the same coding algorithm is used across all the coded speech files. The experiments in this chapter involve passing speech files under investigation through the EVRC and then introducing lost frames in a controlled manner, while keeping all other factors constant.

8.1 Experimental methodology

As shown in Figure 8.1, the speech files were coded at two different anchor operating points (OP0 and OP2), these modes roughly translating into low- and high-quality speech coding. For each mode, the speech files were processed by the EVRC under two scenarios; one assuming no FL and the other with speech frames lost between the encoder and decoder stages of the codec in some controlled manner, as will be discussed next. The rationale behind conducting FVC experiments at two different speech qualities was to try and separate out the impact of speech coding quality from the impact of FL. It is important to note that
8.1 Experimental methodology

Figure 8.1: Processing of speech files using the EVRC at low- and high-quality modes.

the background set used in these experiments contained coded speech at the specific mode being investigated, but without FL.

Figure 8.2 shows a block diagram of the processing stages that were used in these experiments. Again the same FVC experimental procedure as outlined in Section 6.1 has been used for this experiment.

The first aspect that needs to be considered when designing experiments of this kind is what proportion of frames is typically lost or corrupted in an actual network. In the CDMA network the Frame Error Rate (FER) is constantly monitored during a call. When the FER exceeds 10 to 15% it is known that the overall voice quality degrades to a level where the Mean Opinion Score (MOS) is less than about 2.9 [121]. Mobile network operators realise that such voice quality is unpleasant to the listener and therefore they have put procedures in place to drop calls automatically when the FER exceeds 10-15%.

This monitoring of FER would be done over hundreds of frames corresponding to durations of seconds of speech. In the current experiments, however, the vowel segments processed are typically 12 to 15 frames in duration. In comparison to the duration of a vowel segment,
the FER monitoring undertaken by the mobile network operators could be classified as a long-term statistical measure, and there would likely be short periods of time in which the actual FER was much higher. The question then arises as to whether this same upper range for FER of 10 to 15% is also appropriate for much shorter segments such as vowels. In order to answer this question, a number of experiments were conducted where the speech quality of coded vowels is examined using Perceptual Evaluation of Speech Quality (PESQ)\(^1\) [122] for a range of FER values, these are: 0% (i.e., no FL), 10%, 25% and 50%. The PESQ values and their corresponding MOS\(^2\) values have been determined using two input speech samples. The first was 20 frames of speech that contained the word “eight” (the PESQ software requires at least 400 ms of speech to produce an output, which is why the word “eight” was used instead of its corresponding diphthong /aI/). The second input was the

---

\(^1\)PESQ is a tool used to automatically assess the speech quality as experienced by real users in a telephone networks and has been standardised by the ITU-T.

\(^2\)MOS is a speech quality measure, which is typically used in telephony networks to assess the quality of synthesized speech. The MOS scale is obtained through subjective ratings of real users and their views on different qualities of speech[123].
8.1 Experimental methodology

Table 8.1: PESQ and MOS scores for a speech sample coded with OP0 at various percentages of FER.

<table>
<thead>
<tr>
<th>Input 1</th>
<th>Input 2</th>
<th>PESQ</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrupted speech FER=0%</td>
<td>FER=0%  (reference value)</td>
<td>4.50</td>
<td>4.55</td>
</tr>
<tr>
<td></td>
<td>FER=10%</td>
<td>3.58</td>
<td>3.66</td>
</tr>
<tr>
<td></td>
<td>FER=25%</td>
<td>2.33</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>FER=50%</td>
<td>1.97</td>
<td>1.61</td>
</tr>
</tbody>
</table>

The same word coded with OP0, but this time for a range of FER values. The experimental results are shown in Table 8.1.

It is clear from this table that when the percentage of FER is greater than 25%, the speech quality drops to values much lower than the currently used threshold by network operators (MOS=2.9). However, when FER is in the region of 10-15%, it translates into MOS values of the order of 2.9 and higher. This upper range for FER is used in these experiments. Given that the durations of vowel segments were in the range of 12 to 15 frames, the 10-15% FER rate translates into a maximum number of lost frames being typically one, or at most two. For each vowel token analysed, the locations of these lost frames has been determined randomly according to a uniform distribution. The choice of a uniform random distribution here is arbitrary at this stage and further investigation is required as to whether there is a more appropriate distribution that should be used.

It needs to be noted here that the original software implementation of the EVRC has had to be modified in the CDMA-codec platform (see Appendix B) in order to allow specifying a percentage of FL. The total number of lost frames is made equal to the FL percentage specified, multiplied by the number of 20 ms speech frames found in the target speech file and these are selected randomly.
Table 8.2: Impact of FL on FVC performance for the EVRC

<table>
<thead>
<tr>
<th>Mode</th>
<th>$C_{llr}$</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP2 (no FL)</td>
<td>0.116</td>
<td>1.953</td>
</tr>
<tr>
<td>OP2 (with FL)</td>
<td><strong>0.214</strong></td>
<td><strong>1.580</strong></td>
</tr>
<tr>
<td>OP0 (no FL)</td>
<td>0.117</td>
<td>1.892</td>
</tr>
<tr>
<td>OP0 (with FL)</td>
<td>0.161</td>
<td>1.685</td>
</tr>
</tbody>
</table>

### 8.2 FVC performance for two codec modes under FL conditions

Table 8.2 examines the impact of FL on FVC performance in terms of $C_{llr}$ and $CI$ for the EVRC at the two anchor operating points, OP2 (corresponding to low-quality coding) and OP0 (corresponding to high-quality coding). For purposes of comparison, results are presented without and with FL for both cases. A graphical presentation of these results is shown in Figure 8.3.

With respect to accuracy (i.e., $C_{llr}$) of LR results, the FL negatively impacts upon FVC accuracy and this is worse for low-quality speech coding. However, the $CI$ for both low- and high-quality speech coding has actually improved as a result of FL. This unexpected improvement of $CI$ is likely to be a result of the reduced variation across the different recording sessions of each speaker. With EVRC, the loss of a single frame impacts upon subsequent decoded speech frames (see the illustrative example in Section 5.2.2). If for instance two consecutive frames are lost, then speech data for these two and the frame that follows will be derived from the last ‘Good’ frame received. Eventually this will result in four frames that have a similar pattern and thus a reduction in the within-segment variation is expected. This can increase the likelihood of similarity between speech samples coming from the same speaker and, therefore, an improvement in the $CI$ of same-speaker comparisons is likely to be observed (this behaviour is evident in the Tippett plots of Figures 8.5 and 8.7 to be discussed next).

To further understand the degradation in $C_{llr}$ values, Tippett plots have been produced for the two speech coding qualities. Results of the OP0 experiments (without and with FL) are shown in Figures 8.4 and 8.5, respectively. For the OP2 experiments, without and with
FL results are shown in Figures 8.6 and 8.7, respectively. The first observation from these figures is that FL has negatively impacted different-speaker classifications. Secondly, it has caused the same-speaker misclassifications to increase in the case of low speech coding quality (i.e., OP2), whereas it has slightly improved same-speaker classifications for OP0 (almost no misclassifications in this case). This is because with low speech coding quality, a bandwidth expansion on speech data of previous speech frames is performed in order to match the new higher bit rate of the artificially generated frame (i.e., 8.55 kbps). This process, however, is not required for OP0 as the voiced frames under this mode are already encoded at 8.55 kbps. This simple repeat mechanism for lost frames in the case of OP0 reduces the within-segment variation and thus improves same-speaker comparisons.

As far as the CI is concerned, Figures 8.4, 8.6, 8.5 and 8.7 confirm the previous finding, namely FL has caused this aspect to improve, particularly for the same-speaker comparisons.

Figures 8.8 and 8.9 show the APE-plots for modes OP2 (low-quality coding) and OP0 (high-quality coding), respectively. FL in low-quality coded speech causes discrimination loss to increase significantly, in this case by about 110%. There is also a small increase
8.2 FVC performance for two codec modes under FL conditions

Figure 8.4: Tippett plot showing the performance for the OP0 Mode without FL

Figure 8.5: Tippett plot showing the performance for the OP0 Mode with FL
8.2 FVC performance for two codec modes under FL conditions

Figure 8.6: Tippett plot showing the performance for the OP2 Mode without FL

Figure 8.7: Tippett plot showing the performance for the OP2 Mode with FL
in calibration loss of about 30%. The situation for high-quality speech is somewhat different. Here, the major impact of FL is to increase calibration loss by about 65%, with discrimination loss increasing by only about 15%. It is worth mentioning that low speech quality is used in the CDMA network only when the number of users accessing the system is large. This is likely to be the case if a call was made from, say, a busy area such as the centre of a city. Therefore, knowledge of the geographical location of a call may help the forensic speech scientist to decide whether the low- or high-quality speech scenario was more likely.

**8.3 Chapter summary**

Two sets of experiments using two speech coding qualities (low and high) have been conducted to examine the impact of FL on FVC performance. The rationale behind this was to try and separate out as much as possible the impact of speech coding quality (i.e., DRC) and BN from the impact of FL. The percentage of FL chosen for these experiments was in the
8.3 Chapter summary

Figure 8.9: APE-plot showing the performance of FVC using EVRC-coded speech at Mode OP0 without and with FL

region of 10 to 15%. This is the upper range above which the CDMA network drops a call. Given that a single lost frame will also impact upon a number of the subsequent frames that follow it, the amount of artificially generated material in a mobile phone speech recording could well be higher than this range.

Considering the FVC results presented in this chapter, FL negatively impacts upon FVC accuracy and this is worse for low-quality speech coding. The worsening performance is mainly linked to an increase in different-speaker misclassifications. This aspect requires further investigation to determine why this is happening. On the other hand, same-speaker comparisons had actually slightly improved, specifically, in the case of high-quality speech coding. This is because this type of coding uses only the highest bit rate available (8.55 kbps) for coding voiced frames (vowels in this case) and under FL conditions it does not manipulate the data of previous speech frames, but rather it uses a simple repeat mechanism. This reduces the within-segment variation and, in effect, improves the same-speaker comparisons. In terms of reliability, FL improves this aspect in terms of both same- and different-speaker comparisons.
Chapter 9

Impact of Background Noise (BN) at the transmitting end on FVC Analysis

The primary goal of this chapter is to investigate the impact of BN at the transmitting end, in isolation from FL and restricting the change in DRC to fixed AOPs, on the performance of FVC analysis. The most common types of noise in mobile phone communications are babble, car and street noises, and these noise types have therefore been used in these experiments. Details of these experiments are discussed next.

9.1 Experimental methodology

Figure 9.1 illustrates the experimental procedure of this section which involves a comparison in terms of $C_{llr}$ and $CI$, between two FVC analyses. The first is using clean speech that has been processed by the codec at two different anchor operating points: OP0 and OP2. In the second experiment, different kinds of BN, at various Signal-to-Noise Ratio (SNR) levels, were added to the suspect and offender speech files prior to processing them at these same modes of operation (i.e., OP0 and OP2). To reflect more realistic scenarios in these experiments, different sections of noise were added to the speech samples. Therefore, even samples that belong to the same speaker are likely to have different sections of noise being
added to them. Only matched conditions have been considered here, where the background set contains coded speech at the specific mode being investigated, but without BN.

The rationale behind conducting FVC experiments at two different OPs is that OP0 incorporates a different set of coding algorithms than OP1 and OP2. These two use a combination of PPP and CELP for coding vowels, whereas OP0 uses only CELP. The PPP algorithm exploits the fact that pitch patterns do not dramatically change from one frame to another. Thus, rather than transmitting the pitch contour\(^1\) for every speech frame, it uses this information from previous frames while resolving any discontinues that might arise from phase misalignment. The repetition of patterns could mask the effect of Noise Suppression (NS),

\(^1\)The pitch contour is essentially a set of four gains and pitch values calculated every 5ms. A 20 ms speech frame is segmented into four sub-frames of 5 ms each.
which in turn might improve the comparison results. OP1 has not been investigated here because it essentially uses the same coding processes as OP2.

The BN files (car, babble and street noise) have been acquired from the Soundjay database [124]. The SNR levels used in these experiments are typical of those used by the mobile network operators to test the performance of their codecs under various noise conditions and these were: 9, 15, and 21 dB [125].

### 9.2 FVC performance for two codec modes under various BN conditions

Table 9.1 examines the impact of the EVRC on FVC performance under various BN conditions (see Figure 9.1) at the transmitting end. The results are shown in terms of $C_{llr}$ and $CI$. A graphical presentation of these results are also shown in Figures 9.2 and 9.3 for OP0 and OP2, respectively.

Table 9.1: Performance of FVC in the CDMA network using speech coded at OP0 and OP2 under various BN conditions.

<table>
<thead>
<tr>
<th>BN Type</th>
<th>SNR (dB)</th>
<th>OP0 $C_{llr}$</th>
<th>OP0 $CI$</th>
<th>OP2 $C_{llr}$</th>
<th>OP2 $CI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>9</td>
<td>0.185</td>
<td>1.537</td>
<td>0.172</td>
<td>1.793</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.152</td>
<td>1.585</td>
<td>0.152</td>
<td>1.616</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.149</td>
<td>1.655</td>
<td>0.146</td>
<td>1.608</td>
</tr>
<tr>
<td>Babble</td>
<td>9</td>
<td>0.265</td>
<td>1.503</td>
<td>0.190</td>
<td>2.306</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.146</td>
<td>2.262</td>
<td>0.128</td>
<td>2.071</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.144</td>
<td>1.686</td>
<td>0.143</td>
<td>1.911</td>
</tr>
<tr>
<td>Street</td>
<td>9</td>
<td>0.247</td>
<td>1.658</td>
<td>0.209</td>
<td>1.825</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.169</td>
<td>1.645</td>
<td>0.167</td>
<td>1.714</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.143</td>
<td>1.792</td>
<td>0.165</td>
<td>1.914</td>
</tr>
<tr>
<td>Clean-coded speech</td>
<td>0.122</td>
<td>1.861</td>
<td>0.127</td>
<td>2.036</td>
<td></td>
</tr>
</tbody>
</table>

In all experiments, the addition of BN negatively impacts upon the accuracy of LR results. This is worse for high-quality speech coding when the SNR levels are lower than 15 dB.
This is because the task of distinguishing BN from speech at low SNR levels becomes a difficult task for the NS, with removal of parts of the original speech signal likely to result. However, the use of PPP coding in OP2 can somewhat mask this by repeating the pitch pattern of previous frames that might be less distorted. In contrast, at SNR levels around 21 dB, no significant differences have been observed between the two OPs, in terms of both the \( C_{llr} \) and \( CI \). This is because the process of NS is expected to function effectively in both cases at such SNR levels.

The \( CI \) for both speech coding qualities improved for most of the BN experiments compared to the clean-coded speech experiments. This improvement is evident in Figures 9.2 and 9.3, where most of the FVC performance values are located below the reference point (this is the black symbol in Figures 9.2 and 9.3 and refers to the performance of FVC using clean-coded speech). This is an unexpected result and it is not clear at this stage why this is happening.

With respect to differences between the two coding qualities, a negative correlation between the \( CI \) and \( C_{llr} \) values has been observed, when BN levels are high (i.e., with SNRs 15 dB or below). In this case, the \( CI \) increases and \( C_{llr} \) decreases for the lower coding quality as compared to the higher coding quality\(^2\). This behaviour is likely to be linked to the coding patterns used in OP2. As previously discussed in Section 5.2.1, the coding pattern for voiced frames in OP2 is PPP, PPP, CELP, whereas it is always CELP for OP0. This aspect, together with high levels of BN, introduces slightly more variation between the speech samples of all speakers and thus increases the between-session variation. This is evident in the elevated \( CI \) values of the OP2 results at low SNRs. In the case of OP0, the NS subtracts BN from every speech frame without a mechanism, such as PPP, to reduce this effect. This can result in samples that are more similar, but highly distorted, which affects the measurement of speech parameters of interest. As a result, better \( CI \) and worse \( C_{llr} \) can be observed for OP0 as compared to OP2. The situation reverses for these two coding qualities when BN levels decrease (i.e., with SNRs above 15 dB in this case), where the \( C_{llr} \) and \( CI \) values are better for OP0 than OP2, which is an expected behaviour. This again is a result of the NS being able to function effectively at higher SNRs.

\(^2\)It should not be inferred, though, that the accuracy improves as the SNR decreases for low-quality coding (i.e., OP2), but rather the accuracy for OP2 is better at high BN levels as compared to OP0.
In order to examine the BN impact with respect to the same- and different-speaker comparisons, Tippett plots have also been produced for all the experiments with babble noise. Tippett plots for the other types of BN have not been shown here as they are very similar to those for babble noise. Figures 9.4 and 9.5 show Tippett plots for the FVC results using clean speech coded with OP0 and OP2, respectively. Figures 9.6, 9.8 and 9.10 show Tippett plots for the OP0 experiment using speech corrupted with babble noise at 9, 15 and 21 dB SNRs, respectively. Their corresponding OP2 results are shown in Figures 9.7, 9.9 and 9.11, respectively. In the case of high BN levels (9 dB in this case), the addition of BN causes a significant increase in the proportion of both same- and different-speaker misclassifications and these are worse for high-quality speech coding for the reasons previously mentioned. As the SNR increases, it appears that only different-speaker comparisons are negatively impacted. The degree of this impact was almost the same for both coding qualities.

Figure 9.2: $C_{itr}$ v.s. CI for EVRC-coded speech at OP0 under various BN conditions.
Figure 9.4: Tippett plot showing FVC performance of speech uncorrupted with BN and coded at OP0.

Figure 9.3: $C_{llr}$ v.s. $CI$ for EVRC-coded speech at OP2 under various BN conditions.
9.2 FVC performance for two codec modes under various BN conditions

Figure 9.5: Tippett plot showing FVC performance of speech uncorrupted with BN and coded at OP2

Figure 9.6: Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 9 dB coded at OP0.
9.2 FVC performance for two codec modes under various BN conditions

Figure 9.7: Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 9 dB coded at OP2.

Figure 9.8: Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 15 dB coded at OP0.
9.2 FVC performance for two codec modes under various BN conditions

Figure 9.9: Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 15 dB coded at OP2.

Figure 9.10: Tippett plot showing FVC performance of speech corrupted with babble noise at SNR = 21 dB coded at OP0.
In order to analyze the losses in $C_{llr}$, APE plots were produced for the babble noise experiments. Again these were found to be typical of the other types of noise and, therefore, their APE plots are not shown here. The APE plots in Figures 9.12 and 9.13 correspond to the OP0 and OP2 experiments, respectively, using speech files corrupted with babble noise at various SNR levels. Analysis of these plots reveals that the degradation in $C_{llr}$ is mainly attributable to a decrease in the discrimination performance of speech parameters (i.e., $C_{llr_{min}}$). The presence of high levels of BN combined with low-quality coded speech causes the discrimination loss to increase by about 140% in this experiment. The situation was even worse for the higher speech quality, where discrimination loss increased by about 190%. The calibration performance $C_{llr_{cal}}$ for all the cases was found comparable, but it tends to be higher (i.e., worse) for higher-quality speech coding when SNR levels are low.
9.2 FVC performance for two codec modes under various BN conditions

Figure 9.12: APE plots showing the losses in $C_{llr}$ using speech coded at OP0 under various BN conditions.

Figure 9.13: APE plots showing the losses in $C_{llr}$ using speech coded at OP2 under various BN conditions.
9.3 Chapter summary

Two sets of experiments were conducted using various levels of BN added to the speaker’s speech files and then coded using low- and high-quality coding modes: OP0 and OP2, respectively. The FVC results suggest that this aspect of BN at the transmitting end can have quite significant impact on the accuracy of FVC. This is likely to be the case if a call was made from a busy street or near a construction site, etc. This impact is even worse for higher-quality speech coding, which is typically the case if a call was made from a less populated area. This is due to the fact that low-quality speech coding modes, such as OP1 and OP2, can mask the impact of BN by repeating information from previous frames, whereas this mechanism does not occur for high-quality coding (i.e., OP0). The reliability of LR results was not significantly impacted by the addition of BN for either of the investigated codec modes. In fact, this aspect improved for most of the investigated scenarios with BN.
Chapter 10

Overall Impact of the CDMA Network on FVC Analysis

This section examines to what extent the performance of FVC analysis might be negatively impacted by the CDMA mobile phone network as a whole. In order to achieve this, the three major aspects of the CDMA network previously investigated are brought together to reflect more typical situations in the CDMA network. This could provide knowledge of the totality, which is crucially important for forensic speech scientists.

10.1 Experimental setup

Figure 10.1 shows a block diagram of the experimental procedure in this section, which involves a comparison between two FVC analyses in terms of $C_{llr}$ and $CI$. The first FVC analysis is performed using speech that has not been processed by the EVRC, and referred to here as the clean speech experiment. These results are the same as those presented in Section 6.2. The second set of FVC analyses uses speech that has been processed by the EVRC, taking into account three key aspects of the CDMA network: (i) DRC (i.e., capacity conditions), (ii) FL (i.e., channel conditions) and (iii) the presence of BN at the transmitting end. One could actually conduct a large number of experiments given all the
10.1 Experimental setup

Figure 10.1: Block diagram showing the experimental procedure to assess the overall impact of the CDMA mobile phone network on FVC.

Possible operational modes of the codec under various conditions in the network. However, only typical and representative scenarios have been considered here.

In terms of DRC, Medium-Quality (MQ) speech coding has been used. This is the situation when a cellular site is neither congested nor has only few users (i.e., medium channel capacity). With MQ speech, the EVRC switches between all the three anchor operating points. This is particularly important, as mentioned previously (see Chapter 9), as OP1 and OP2 use different algorithms for coding voiced speech and these can sometimes mask the damage caused by NS when compared to OP0. As with the DRC experiments in Section 5.2.1, discrete ADR values were used for coding the speech files to ensure a better coverage of the selected range (in this case the ADR values can take on any value from 4.8 to 9.6 kbps). Using the CDMA codec platform, these values were randomly selected for the
speech files using the medium capacity ADR group (4.8, 5.8, 6.2, 6.6, 7.0, 7.5, 8.5, 9.6 kbps) following a uniform distribution.

Given that the wireless channel in mobile phone networks is relatively poor, there is a high possibility of frames being lost or corrupted during transmission. In order to incorporate this situation in the current investigation, the maximum upper range of FL (10-15%) has been considered and the temporal locations of lost frames have been selected randomly for the target speech files. It needs to be noted here that FL percentages less than this would translate into a half or quarter lost frame, which would make no sense, as partially corrupted frames cause the whole frame to be lost. Given the short segments of speech used in these experiments, one lost frame translates to 10-15% FL, depending on the duration of the speech file.

The third major aspect which must be considered in this kind of experiment is the presence of BN at the transmitting end. Mobile speech recordings are often contaminated with BN picked up by the microphone at the transmitting end. In order to reflect the impact of this aspect as well, babble noise was added to the speech files prior to processing them with the EVRC, and this was at three different SNRs: 9, 15, and 21 dB. With respect to the background set, this contained only clean uncoded speech in the case of the uncoded speech experiment.

One type of mismatch, namely studio-quality and mobile-phone-quality recordings mismatch, and its respective impact on FVC has been briefly examined in this chapter. Two sets of experiments have been conducted, the first one using MQ-coded speech in the background set (i.e., matched conditions) and another using clean, uncoded speech (i.e., mismatched conditions). The aim with these two being to examine the extent to which mismatch conditions might impact upon the accuracy and reliability of LR results. Table 10.1 shows how the suspect, offender and background samples have been coded under matched and mismatched conditions.
Table 10.1: Transmission scenarios for the suspect, offender and background samples under matched and mismatched conditions.

<table>
<thead>
<tr>
<th>Data</th>
<th>Matched</th>
<th>Mismatched</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DRC</td>
<td>FL</td>
</tr>
<tr>
<td>Offender</td>
<td>MQ</td>
<td>10-15%</td>
</tr>
<tr>
<td>Suspect</td>
<td>MQ</td>
<td>10-15%</td>
</tr>
<tr>
<td>Background</td>
<td>HQ</td>
<td>Null</td>
</tr>
</tbody>
</table>

10.2 FVC Performance under matched/mismatched conditions

10.2.1 Matched

Table 10.2 compares the resulting FVC performance for a number of typical scenarios in the CDMA network. The uncoded speech results are also shown. To facilitate a better comparison, a graphical presentation of these results is shown in Figure 10.2. On the basis of $C_{llr}$ results, it can be noted first that the CDMA-coded speech always resulted in a negative impact on the accuracy of the FVC results when compared to uncoded speech. Second, the $C_{llr}$ results improved with increasing SNR levels, which is an expected trend. A comparison with respect to the uncoded speech experiment indicates that this increase in $C_{llr}$ can be as high as 70%.

In contrast, the precision of LR results has actually improved for all the investigated scenarios. The 15 dB experiment might seem like an outlier, but this is not actually the case. In these experiments, lost frames are selected randomly and therefore the temporal location of lost frames may differ from one recording session to another (even between speech files for the same speaker). As previously noted in Chapter 9, speech coding modes, such as OP1 and OP2 (which are used for MQ speech), can mask noise suppression damage by repeating the pitch pattern of previous frames. Frame erasure in the EVRC also inherently repeats information from previous frames and when these two situations (i.e., BN and FL) are combined, the EVRC decoder begins to repeat frames that have been affected by noise suppression, depending on the temporal location of the FL. This might increase the
Table 10.2: Performance of FVC in the CDMA network using medium-quality (MQ) speech coding, 10-15% FL under various BN levels with matched conditions.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>SNR (dB) of Babble Noise</th>
<th>$C_{llr}$</th>
<th>$CI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL=10-15%, MQ Speech Coding</td>
<td>9</td>
<td>0.279</td>
<td>1.681</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.249</td>
<td>2.157</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.181</td>
<td>1.648</td>
</tr>
<tr>
<td>Uncoded (or clean) Speech</td>
<td></td>
<td>0.167</td>
<td>2.299</td>
</tr>
</tbody>
</table>

between-session variation and thus worsen the reliability of LR results, as observed for the 15 dB scenario. Nonetheless, in all scenarios, this worsening $CI$ performance is still better than for the uncoded speech. Again, this is attributed to the quantisation process inherent in the EVRC which reduces the degree of variability between speech samples.

To further examine the FVC results, Tippett plots have been produced for the investigated scenarios (9, 15 and 21 dB) as shown in Figures 10.4, 10.5 and 10.6, respectively. The Tippett plot in Figure 10.3 shows FVC performance for the uncoded speech experiment (details about this experiment can be found in Section 6.2). It is clear from Figures 10.4, 10.5 and 10.6 that the proportions of both same- and different-speaker misclassifications are higher for coded speech than uncoded. As the SNR level decreases, it is mainly the same-speaker comparisons that worsen.

The losses in $C_{llr}$ were further examined using a set of APE plots produced for the investigated scenarios, as shown in Figure 10.7. In all cases, including the clean speech experiment, $C_{llr}$ is dominated by the discrimination loss $C_{llr_{min}}$. The relative increase in $C_{llr_{min}}$, as compared to the uncoded speech results, varies from 10 to 90% depending on the SNR level, with 21 dB having the least impact and 9 dB being the greatest. In terms of calibration loss $C_{llr_{cal}}$, no significant differences were observed between the uncoded speech results and the three investigated scenarios under matched conditions.
10.2 FVC Performance under matched/mismatched conditions

Figure 10.2: A bar graph showing the FVC performance ($C_{llr}$ and $CI$) for different scenarios in the CDMA mobile phone network under matched conditions.

Figure 10.3: Tippett plot showing FVC performance of uncoded speech using.
10.2 FVC Performance under matched/mismatched conditions

Figure 10.4: Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 9 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).

Figure 10.5: Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 15 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).
Figure 10.6: Tippett plot showing FVC performance of medium speech coding quality. The original speech files are corrupted with babble noise at SNR = 21 dB and 10-15% FL. The background set contains MQ-coded speech (i.e., matched scenario).

Figure 10.7: APE plots showing losses in $C_{llr}$ using speech coded under matched conditions.
10.2 FVC Performance under matched/mismatched conditions

10.2.2 Mismatched

Table 10.3 compares the resulting FVC performance for the same scenarios indicated in the previous section, but this time using uncoded speech for the background set. A graphical presentation of these results is shown in Figure 10.8. In terms of $C_{llr}$, the results of mismatched conditions are substantially worse than for matched conditions. A comparison with respect to the uncoded speech experiment shows an increase in the $C_{llr}$ value by about 100% to 200% depending on the SNR level. The accuracy of the LR results under mismatched conditions are approximately 130% worse than matched conditions at $SNR = 9\, dB$. With respect to the reliability of LR results, this has actually improved, which is an unexpected result. Why this might be so is not clear at this stage. However, inaccurate but precise results are clearly undesirable. Therefore, it is advised that practitioners use coded speech for the background set in order to minimize the impact of this aspect.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Codec Settings</th>
<th>SNR (dB) of Babble Noise</th>
<th>$C_{llr}$</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FL=10-15%, MQ Speech Coding</td>
<td>9</td>
<td>0.506</td>
<td>0.855</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.358</td>
<td>1.396</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21</td>
<td>0.343</td>
<td>1.201</td>
</tr>
<tr>
<td></td>
<td>Uncoded (or clean) Speech</td>
<td></td>
<td>0.167</td>
<td>2.299</td>
</tr>
</tbody>
</table>

The FVC results were again investigated by looking at their respective Tippett plots. These have been produced for the three previously investigated BN scenarios (9, 15 and 21 dB) as shown in Figures 10.9, 10.10 and 10.11, respectively. It is clear from these figures that the proportions of same- and different-speaker misclassifications are significantly higher than for both the uncoded and matched condition scenarios. This trend is observed to be consistent across all three scenarios.

As with the matched condition experiments, the losses in $C_{llr}$ were further examined using APE plots (Figure 10.12). $C_{llr}$ is again dominated by the discrimination loss $C_{llr_{min}}$, but the increase in $C_{llr_{min}}$ is quite significant this time. A comparison with the uncoded speech experiment shows that $C_{llr_{min}}$ increased by about 100 to 180%, depending on the SNR level,
10.2 FVC Performance under matched/mismatched conditions

Figure 10.8: A bar graph showing FVC performance ($C_{llr}$ and $CI$) for different scenarios in the CDMA mobile phone network under mismatched conditions.

Figure 10.9: Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 9 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario).
10.2 FVC Performance under matched/mismatched conditions

Figure 10.10: Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 15 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario).

Figure 10.11: Tippett plot showing FVC performance of medium speech coding quality. The original speech files were corrupted with babble noise at SNR = 21 dB and 10-15% FL. The background set contains uncoded speech (i.e., mismatched scenario).
which varied in this case between 21 and 9 dB, respectively. There is a substantial increase in the calibration loss $C_{llr_{cal}}$ as well, specifically, when mismatched conditions are accompanied with high levels of BN (9 dB in this case). The increase in $C_{llr_{cal}}$ can be as high as 270%, which indicates that the FVC system begins to lose its computational performance. Whether this might be the case for other FVC systems, such as the GMM-UBM, requires further investigation. These results, nonetheless, confirm the previous conclusion, namely that using coded speech instead of uncoded speech for the background set can substantially improve the accuracy of LR results when dealing with CDMA mobile phone speech data. However, whether using speech recordings in the background set acquired from a different network origin, such as the GSM network, might have the same impact on FVC results is an aspect that also requires further investigation.
10.3 Comparison between impacts of the GSM and CDMA networks on FVC

A study of the GSM network and its impact on FVC by Nair [89] has been recently published by the University of Auckland. It was felt necessary and useful to practitioners at this stage to include a brief comparison between both networks.

With respect to DRC and its impact on FVC in the GSM network, it was shown in [89] that in virtually all the cases of DRC, both the accuracy and reliability of FVC analysis are better for coded speech than uncoded. This aligns with the findings of this study, namely the DRC aspect improves the performance of FVC. Experiments with different speech qualities in the GSM network did not show any consistent trend between FVC accuracy and speech quality. For CDMA, however, the accuracy improves as the speech quality improves. It needs to be noted here, that the driving force for changing speech quality in the GSM network is changing channel conditions, whereas in the CDMA network it is mainly a function of the number of users accessing the system (i.e., number of users present in a cell site).

The FL aspect was studied in isolation from DRC and BN in the GSM network, whereas in the CDMA network DRC can never be completely disabled (see Chapter 5). FL in the GSM and CDMA networks was found to have quite a significant impact on FVC results. The worsening performance of GSM-coded speech in the case of FL was mainly due to increases in same-speaker misclassifications. In the CDMA experiments, FL mainly affected the different-speaker comparisons. In both networks, the situation is worse when FL is accompanied with low-quality speech coding. It was found in [89] that discrimination loss is the main cause for the poor accuracy of FVC, irrespective of the speech-coding quality being used. For CDMA-coded speech, this was only the case for low-quality speech. With high-quality coding, the degradation in accuracy was mainly due to calibration losses. Upon FL, the reliability was slightly negatively impacted in the GSM experiments. In contrast, this aspect improved for CDMA-coded speech.

No strategies have been put in place in the GSM network to reduce the impact of BN at the transmitting end. It was found in Nair work [89] that this aspect negatively impacts upon the
accuracy of FVC results, as compared to clean-coded speech. This same finding applies to
CDMA-coded speech. The accuracy of FVC becomes worse as the speech-coding quality
reduces in the GSM network. The opposite was found, in this study, when using CDMA-
coded speech. The accuracy was better for low-quality than high-quality speech. This is
because low-quality coding in the EVRC incorporates a mechanism that sometimes mask
the damage caused by BN. In terms of the reliability of FVC, this has slightly improved for
GSM-coded speech upon adding BN. The improvement in reliability was more noticeable
in the CDMA results.

With respect to the overall impact of DRC, FL and BN on FVC, the integration of these as-
pects was found to worsen the accuracy of FVC in both networks. The reliability improved
in both the cases. One mismatch scenario was investigated in both networks using clean
speech in the background set. The GSM results suggest that the impact of such mismatch
scenarios can marginally worsen the accuracy of FVC compared to using coded speech in
the background set (i.e., matched conditions). On the other hand, it is highly recommended
to use coded speech in the background set when dealing with CDMA-quality speech. Fail-
ure to do this may result in a significantly worse accuracy. In both networks, the use of
clean speech in the background set has in fact improved the reliability of FVC analysis.
However, inaccurate but precise is clearly undesirable.

10.4 Chapter summary

The integration of all three aspects (DRC, FL and BN) was considered in this chapter to
investigate the impact of more realistic conditions in the CDMA network on FVC. Though
the investigation in this chapter was restricted to only three representative scenarios, other
scenarios that were previously discussed in Chapters 7, 8 and 9 could also happen in reality.
For example, there is a possibility that only the DRC aspect is occurring, but the call does
not suffer from any worsening BN or FL conditions. The results of such scenarios are
presented in Chapter 7.

With respect to the experiments of this chapter, an average number of users was assumed to
be present in a cell site, and given that DRC is a function of capacity, the resultant speech
translates into MQ. This was used for all the investigated scenarios. FL is assumed to be at its upper range of 10-15%. The reason for not conducting experiments at lower percentages of FL is that short segments of speech have been used. The 10-15% FL would normally translate into one or a maximum of two frames and so percentages lower than this were not meaningful. Situations where there is no FL have already been investigated in Chapters 7 and 9, when aspects such the DRC and BN at the transmitting end were examined in isolation from FL.

FVC results of the investigated scenarios show that both the same- and different-speaker comparisons were negatively impacted. The situation became even worse for same-speaker comparisons when BN levels are high. The reliability of LR results was not significantly impacted when using mobile phone speech and, in fact, slightly improved. It can also be concluded that using uncoded (or clean) speech for the background set could significantly worsen the comparison results. Hence, practitioners are strongly advised to use only coded speech for the background set. Fortunately, the EVRC-codec platform developed as part of this research facilitates the transformation of existing speech files into CDMA-quality speech, including those of the background set, without the need to transmit speech files through an actual network. This procedure was found to be less time consuming and more representative of the actual possible scenarios in the network.
Chapter 11

Conclusions and Directions for Future Work

11.1 Conclusions

In this study the impacts of three major aspects of the CDMA mobile phone network on the strength of speech evidence and the performance of FVC were investigated. From the outset, forensic speech scientists need to appreciate that within the mobile phone arena there are different technologies such as GSM and CDMA and these incorporate different mechanisms in respect to how they handle the speech signal. Therefore it is to be expected that the impact of these networks on the speech parameters of interest in FVC is going to be different. The three key aspects of the CDMA network that can negatively impact the speech signal are DRC, the FL mechanism and handling of BN. A thorough understanding of these has been used in this research to drive the EVRC in a manner similar to an actual CDMA mobile phone network.

A dedicated software platform was developed to simulate realistic scenarios in the CDMA network and examine the impact of each of the above aspects on speech, either separately or in combinations. Given that in the CDMA network it is the speech codec alone that directly controls and impacts speech quality, these experiments were conducted by focusing solely on all possible modes of the EVRC. This strategy makes it possible to encompass a large
number of possible scenarios in the CDMA network. This argued to be a much better approach than undertaking numerous experiments involving transmission of speech over an actual mobile phone network. There would have then been no way of knowing what possible transmission scenarios had been represented, as such information is not available in the received speech signal. This software platform can also assist forensic practitioners to transform their existing databases to EVRC-coded speech when conducting real world FVC analysis. Further, it allows investigating the impact of DRC in isolation from FL and BN at the transmitting end. Similarly the impact of FL can be studied in isolation from BN and vice versa, while restricting the change in DRC to one of the AOPs. DRC can never be completely disabled in EVRC, even if no changes in channel capacity occur, because it constantly produces speech frames at different bit rates depending on their classification, such as: voiced, unvoiced or transient. However, by restricting the codec to a specific AOP, the coding algorithms used then becomes fixed.

A new approach was presented in this thesis for computing likelihood ratios, termed Principal Component Analysis Kernel Likelihood Ratio (PCAKLR). The primary motivation for this approach was the mathematical robustness issues found with one of the popular approaches for computing LRs, namely MVKD. The cause of fragility in the MVKD algorithm is linked to the inversion of ill-conditioned matrices as well as the difficulty of smoothing the kernel density estimation when the number of input parameters is somewhat larger than three or four. In contrast, the PCAKLR approach does not require the inversion of any matrices, and handles one input parameter at a time, which makes it computationally robust. The idea behind PCAKLR is simple. A set of input parameters is first transformed into a new set parameters with little to no correlation using principal component analysis (PCA). LRs are then computed using the univariate kernel density analysis (UKD) for each transformed parameter. The product of these individual LRs is then taken to produce an overall LR based on the naïve Bayesian approach.

PCAKLR was found to exhibit a very similar performance to MVKD when used with a small number of input parameters, in the region of three or four. As expected, the performance of PCAKLR improves when the number of input parameters is increased. This approach also has an interesting feature which permits combining LR results for individual
speech segments without using the conventional method, namely logistic-regression fusion. Given PCA does not distinguish between within-segment or between-segment correlations and can extract information from large confusing data sets, the parameters from all individual segments can be merged together into one superset. PCA is then applied to transform this into a new set of orthogonal parameters. The UKD analysis can then be used to process each of these parameters and the resulting LRs combined using simple multiplication. This procedure is referred to as PCAKLR\textsubscript{NF}. This alternative fusion strategy to logistic regression is found to produce marginally more accurate results and has been shown to marginally outperform MVKD for small numbers of parameters. But, as for other LR analysis techniques, the LR results still need to be calibrated using a development set. Though PCAKLR\textsubscript{NF} is a viable alternative for fusing the LR results, the traditional approach (i.e., logistic regression) was used throughout this study.

A number of experiments were conducted to decide upon the best performing parameter set in the context of FVC when dealing with CDMA speech recordings. To do this, the FVC performance of four types of cepstral coefficients (CCCs, RCCs, LPCCs and MFCCs) and formant trajectories were examined for uncoded and coded speech. In the case of uncoded speech (i.e., clean speech), CCCs outperformed all other types of parameters in terms of both accuracy and reliability of LR results. This is because CCCs carry more information about glottal shaping and lip radiation filters. However, this did not hold true for the CDMA experiments as the speech coding process ignores information about these filters, which, from the perspective of perceptual quality, only add a slight improvement. As MFCCs do not contain information about specific parts of the source-filter model, they outperformed the other types of speech parameters.

In respect to impact of the CDMA network on FVC, the first aspect investigated was DRC. This can result in inconsistent quality of transmitted speech over short intervals of time, namely 20 ms. The primary driver of this process is to mitigate the impact of changing channel capacity. Increases in the number of users accessing the system (i.e., higher channel capacity) can increase interference levels between users sharing the same channel.
Conclusions

resources. Reducing the speech coding bit rate can reduce co-user interference. Three different channel capacity conditions were investigated, namely high, medium and low, which in turn translate into low-, medium- and high-quality coded speech, respectively.

One surprising result from the DRC experiments was that in virtually all cases, both the accuracy ($C_{llr}$) and reliability ($CI$) of LR results were better for coded speech than for uncoded. This is likely to be linked to the parameter quantisation processes inherent in the EVRC which reduce intra-speaker variability. This then primarily improves same-speaker comparisons. The results also show a clear link between the quality of the coded speech and $C_{llr}$, with high-quality coded speech having the best FVC accuracy. But the corresponding situation in respect to $CI$ is unexpected, with low-quality speech having a marginally better performance in this regard than either medium- or high-quality speech. The improvement in $C_{llr}$ was further examined in terms of the discrimination and calibration losses using APE–plots. For EVRC-coded speech, an increase in speech quality resulted in a reduction in discrimination loss compared to uncoded speech, but there was no such consistent trend as far as calibration loss is concerned.

In terms of lost or corrupted frames, the CDMA network replaces those with synthetically generated frames based upon past ‘Good’ speech frames, a process implemented by the decoder section of the EVRC. The FL mechanism implemented by the EVRC has been described in considerable detail. It is essential for all forensic speech scientists to have a clear appreciation of the nature and extent to which speech acquired from the CDMA mobile phone network could contain artificially generated sections, a fact that must necessarily impact on the confidence they ascribe to their analysis results. It has been noted that mobile phone network operators prevent a call from continuing if the percentage of lost frames exceeds in the region of 10 to 15%. However, given that a single lost frame will impact upon a number of the ‘Good’ subsequent frames that follow it, the amount of artificially generated material in a mobile phone speech recording could be higher than this.

In the FL experiments, a worst-case scenario was considered using the maximum upper range of FL allowed in a mobile phone call (i.e., 10-15%). Given that the duration of processed speech segments in this investigation was in the range of 10 to 15 frames, this translates into one or two lost frames per segment. In this study, the location of these has
been determined randomly in accordance with a uniform distribution. In the absence of any information about the nature of this distribution in reality, this was considered the most reasonable choice. FL negatively impacts on the accuracy of both same-speaker and different-speaker comparisons. However, in terms of reliability, FL actually improves both of these. FL, together with low-quality speech coding, causes discrimination loss to increase significantly, especially for different-speaker comparisons. The situation for higher-quality speech is somewhat better, where same-speaker comparisons are slightly improved and the $C_{ll}_{h}$ is found to be marginally worse than cases without FL. This improvement is linked to the fact that EVRC does not perform bandwidth expansion on the speech data of previous voiced frames that are coded with OP0\(^1\). This reduces the within-segment variation and as a result improves the same-speaker comparisons. APE plots for high-quality coded speech suggest that the major cause of the impact of FL was, surprisingly, calibration loss and not discrimination loss. In terms of $CI$, this parameter improved in all cases of FL. This is due to the fact that the FL mechanism utilizes information from previous frames, such as the pitch or LPC coefficients, when replacing the lost frames, and often the one following it. This reduces within-segment variation and increases the likelihood of similarities being formed between speech samples coming from the same speaker. Therefore, an improved $CI$ is observed.

In respect to the impact of BN, the process of Noise Suppression (NS) is used in the EVRC to subtract BN from each input speech frame, improving the quality of coded speech. However, the downside of this is that it might cause distortion to the original speech signal when BN levels are high. As expected, the presence of BN in the EVRC-coded speech had a negative impact upon the accuracy of FVC results and this was worse for higher levels of BN. What was unexpected was that low-quality speech coding can result in better accuracy than high-quality speech coding if the SNR levels are low (less than 15 dB). This is because low-quality speech coding, which manifests itself in anchor operating points OP1 and OP2, can mask the damage caused by NS by repeating information from previous frames, a process called PPP. Another unexpected result was the improvement in $CI$, which was observed irrespective of the speech coding quality being used. Why this might be so requires further investigation.

\(^1\)Quite often this is the mode used for high-quality speech coding.
In order to simulate more realistic scenarios in the CDMA network, another set of experiments was conducted taking into account impacts of all three aspects combined. In these experiments, the selected speech coding quality was medium, on the assumption that calls were made from a site in which an average number of users were present (i.e., medium capacity). Another reason for choosing medium-quality speech was the fact that it forces the EVRC to switch between all three anchor operating points. OP1 and OP2 use different algorithms for coding voiced speech to OP0, which can sometimes mask the damage caused by NS. Hence, it was interesting to examine the impact of all the three modes at one time.

In respect to FL, given that the target speech segments were only 10 to 15 frames long, an attempt to introduce even one or two lost frames can translate into about 10-15% of FL. This range matches the maximum acceptable percentage of FL in the CDMA network and thus it was used in these experiments. The BN and noise suppression had little impact on the speech signal in the case of low BN levels, but this impact increased as the levels of noise increased. Hence, BN was added to the speech files at various SNR levels.

The experiments have been conducted twice using coded (matched conditions) and uncoded (mismatched conditions) speech in the background set. A comparison between the results suggests that mismatch is likely to have a negative impact on the accuracy of a FVC analysis. In terms of the accuracy of LR results, the impact was significantly worse under mismatched conditions compared to matched. There were no issues with reliability of FVC analysis under mismatched conditions. In fact, this parameter improved in most cases, most likely, for the same reasons previously discussed.

One important finding from this research is that forensic speech scientists need to be careful when conducting FVC analysis using mobile phone speech. Knowledge about the location (or where the call was made from) can be very useful to determine the possible scenario and the amount of degradation that might have occurred. If for example a call was made from a quiet place (e.g., say a house in a suburb) and the auditory analysis confirms that the amount of BN present in this call was small, by examining silence frames, then it is likely that both the accuracy and reliability of FVC analysis will improve as compared to...
 uncoded speech\textsuperscript{2}. The improvement in the first is likely to be marginal, but the latter could be significant (i.e., results are more reliable). On the other hand, if a call was made from a busy area or street in a city centre and a relatively large amount of BN was found in this call, then the accuracy will be worse than for uncoded speech. Nonetheless, the reliability is not likely to be affected. However, if this same call was made from a less populated area (e.g., say a small factory in a remote area, where high levels of BN are likely to be present), then the accuracy is going to be worse than if this same call was made from a city centre. Moreover, the further the caller is from the base station and the more users present (high co-user’s interference) the more likely FL will occur. Determining from the recovered speech signal when this process has occurred is likely to be very challenging, if not impossible. In worst case scenarios, the speech codec can insert up to 320 ms of artificial speech, which then makes the LR results of any FVC analysis highly questionable.

11.2 Recommendations to forensic speech scientists

In real life cases, forensic samples to be compared are often recorded through different channels. Quite commonly, the crime scene recording is telephone or mobile phone speech and the other is a direct recording through a microphone. A number of studies have shown that reducing channel mismatch can improve the results of FVC analysis Nair et al. [28], Alexander [29], Zhang et al. [27]. However, the optimal procedures used for telephone speech cannot be applied to mobile phone speech. With telephone recordings, the speech signal is transmitted with high fidelity at relatively high bit rate (64kbps) as compared to mobile phone speech, which is highly synthetic Guillemin and Watson [4], Alzqhoul et al. [1]. Given landline networks band limit the speech signal to approximately 4 kHz, then low-pass filter at 4 KHz can also be applied to the studio quality recordings. This is because any information related to the high order formants is likely to be lost in telephone speech. Compensation techniques can then be applied to the offender telephone recordings in order to remove speech artefacts caused by medium of transmission Alexander [29], Kim

\textsuperscript{2}This conclusion is valid on the assumption that the suspect and background data are coded (i.e. matched conditions) in accordance with the above scenario, which is high-quality speech coding in this case (see Chapter 7).
and Hansen [22], Morales et al. [23]. In contrast, the wireless channel in mobile phone networks does not have a direct impact on speech. Factors such as interference, channel fading and channel noise do not impact the speech signal directly, but rather indirectly. As previously discussed in Chapter 5, the speech codec will be instructed by the network to mitigate the impact of these factors. Channel noise, for example, can result in frames being partially corrupted or completely lost. In both cases, the codec will either correct the impacted frames or activate the FL mechanism to replace them. The resultant speech is always clean and thus channel noise will never be directly added to speech. The best advice that can be given at this stage to practitioners is to process the suspect and background data using the CDMA codec. Fortunately, the platform developed in this study (see Appendix B) makes it relatively easy to do this and practitioners are not required to understand the specifics of the underlying mechanisms used in the CDMA network.

Obviously, the next question is: Under which scenario in the CDMA network, the speech files must be processed?. There is a continuum range of possibilities for coding, but the answer to this question can be one of two options. The straightforward option is to choose the middle way. One can assume MQ speech coding and moderate levels of FL and BN. The more accurate way, however, depends if information is available about the geographical location of the call (i.e., where the call was made from). If a call was made from a quite house in an urban area then probably high-quality coding with little BN noise is a viable option to consider. If a call was made from a city centre, then low-quality coding with moderate levels of BN and no FL can be assumed. Another possible mismatch condition is the case where suspect and offender recordings are acquired from different mobile phone networks. The impact of this scenario on Forensic Voice Comparison (FVC) has been investigated by Balamurali Nair and I and results were published in Nair et al. [28]. It was found in this study that the resulting degradation on the accuracy of FVC analysis in such cases can be very significant, but the reliability may actually improve. It was shown that the impact of mismatch conditions can be lessened by

---

3 The choice of no FL here is made on the assumption that the coverage of mobile phone networks at city centres is expected to be reasonably good, and therefore, users are likely to be in close distance to the base stations. In such cases, the SNR of the transmission channel is likely to be good, making the loss of frames unlikely.
passing the suspect speech through the GSM or CDMA codecs, depending on the network origin of the offender’s data. In respect to background data, the optimum FVC analysis performance was found when processing the background speech files twice, first using the GSM codec (AMR) and second using the EVRC. Though this goes a long way to mitigating the impact of mismatch, but it was still not as good as result of matched conditions.

The results in Chapter 9 suggest that when moderate to high levels of BN are present, low-quality coding provides better FVC results than high-quality coding. Preliminary experiments in Section 5.2.3, show that low-quality coding can mask the damage caused by NS to some extent. This is because OP2 uses two consecutive PPP frames for coding voiced speech (for further details see Section 5.2.1). The PPP algorithm repeats the pitch information from previous frame and thus can mask the impact of NS (see Figure 5.9). Perhaps, one can exploit this feature and process the CDMA recordings with low-quality coding to improve the performance of FVC. This is one aspect that needs to be further examined.

Another aspect of interest is the fact that FVC analysis performs better with EVRC-coded speech than uncoded (or cleans speech) and the higher the coding quality the better the performance (see Chapter 7). This may suggest processing the clean speech files with EVRC before conducting FVC analysis, which is not unreasonable thing to do. However, one needs to be careful here, because this recommendation is subject to using MFCCs in the analysis. It has been shown in Chapter 6 that CCCs actually perform better with clean speech than using MFCCs, as they essentially carry more speaker specific information.

It needs to be emphasized here that the damage caused by FL and DRC cannot be reversed. This is because the underlying bit rates used for coding or information about the channel conditions are not available in the received speech signal. As mentioned earlier in Section 5.2.2, determining whether FL occurred or not is a very challenging task. However, it might be possible in the future to develop a neural network and train it using a large number of CDMA samples coded under different transmission scenarios. Then use the trained model to predict the underlying coding and channel conditions. Again this another aspect that requires further investigation.
11.3 Directions for future work

There are several aspects of this research which are still open for investigation. Given differences between the GSM and CDMA technologies, the extent to which these might impact upon the outcome of FVC is expected to be different. It would, therefore, be interesting to have a comparison between FVC results for both networks using speech transmitted under similar conditions. Another important aspect here, which plays a role in improving the accuracy FVC results, is the choice of coded or uncoded speech for the background set. The results in this study suggest that coded speech can significantly improve the FVC results when dealing with CDMA coded speech, but this may not necessarily be the case with GSM mobile phone speech.

Moreover, it is important to investigate the impact of various mismatch conditions between the recordings of offender, suspect and relevant background population. This includes using different sample sizes (i.e., the number of tokens available for each speaker), different dialects (e.g. using a mixture of dialects for the background set that are different to the suspect and offender dialect) and different transmission channels. It would also be interesting to investigate the impact of mismatch between GSM and CDMA mobile phone recordings (e.g. suspect recording being acquired from the GSM network and offender recording from the CDMA network). In such cases, if one decides that coded speech must be used for the background set, it would raise another concern as to whether this should be GSM or CDMA coded, or even a mix of both. An alternative strategy is coding speech files of the background set twice, using the GSM codec first and then the CDMA codec.

The impact of the CDMA network on FVC needs to be further examined using the GMM-UBM model, which is one of the widely used and accepted procedures in this arena, and then compare these results and findings to those of PCAKLR. Unfortunately, given the limited time available for this study and the necessity of having a larger amount of data for conducting experiments with the GMM-UBM, this approach was not considered in these experiments.

Cepstral coefficients are generally known to be sensitive to transmission artefacts in landline networks and various techniques have been proposed to compensate for this. In mobile
phone networks, however, the manner in which the transmission channel impacts upon the speech signal is different and this impact is rather indirect. For example, channel noise, fading, interference and changes in the number of users accessing the system are all resolved in the CDMA network by means of instructions sent to the EVRC to change its mode of operation and counteract these effects. Therefore, developing an alternative strategy for compensation, especially designed for mobile phone speech, is an important area for future research.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


[70] P. Rose, “Catching criminals by their voice–combining automatic and traditional methods for optimum performance in forensic speaker identification,” Australian National University · School of Language Studies, Canberra, Australia.

REFERENCES


REFERENCES


REFERENCES


[107] 3GPP, “Ts 26.071 v11.0 3rd generation partnership project; technical specification group services and system aspects; mandatory speech codec speech processing functions; AMR speech codec; general description.” Retrieved June 10, 2013, from http://www.3gpp.org/, copyright 2012.


REFERENCES


REFERENCES


Appendix A

Speech Parameters for Comparison

This appendix overviews different types of speech parameters used in this thesis for investigating the impact of the CDMA mobile phone network on FVC. The speech parameters discussed here are: (i) Complex Cepstral Coefficients (CCCs), (ii) Real Cepstral Coefficient (RCCs), (iii) Mel-Frequency Cepstral Coefficients (MFCCs), (iv) Linear Prediction Cepstral Coefficients (LPCCs), and (v) Formant Trajectories (FTs). Each of these parameter sets carries information about specific parts of the speech production model (discussed in the first section of this appendix). This background information gives insight as to why the discrimination performance of certain parameters might be superior to others when dealing with CDMA mobile phone speech.

A.1 Speech production and the source-filter model

The source-filter model is one of the widely accepted models with respect to the physical production of sounds [126, 127]. The interaction between different parts of this model is mathematically viewed as a convolution between a number of systems (see Figure A.1). In speech production theory, there are two main sources of energy (i.e., normally referred to as excitation) depending on the type of speech being produced: voiced or unvoiced. In voiced speech the energy comes from air expelled from the lungs into the vocal folds which causes them to open and close at a certain rate, which is called the fundamental frequency (F0).
The motion of the vocal folds modulates the air stream and produces acoustic energy having a spectrum of harmonics at multiples of F0. The air flow builds up slowly in the opening phase, but decreases faster in the closure phase [128]. This is why glottal pulses generally appear to have steeper slope in the closure phase. As a result, the excitation source can be modelled as a train of impulses, at a rate of F0, which then passes through a glottal shaping filter [129]. The effect of spectral slope (i.e., the opening and closure phase slopes) of the excitation source is reflected in the glottal shaping filter. In the case of unvoiced speech, the vocal folds are held apart producing irregular air flow. Fourier analysis of the resulting acoustic energy has shown that the excitation source is a mix of many frequencies and thus is normally modelled as noise. The filter part of the source-filter model comprises a set of three cascaded filters (as shown in Figure A.1): (i) the glottal shaping filter, (ii) vocal tract filter (which relates to the geometry of the vocal tract), and (iii) the lip radiation filter [39, 40].

In order to understand the concept of cepstral analysis and relate cepstral coefficients to the different parts of the source-filter model (discussed in the following sections), it is important to overview the mathematical representation of the speech model. Speech is a result of convolution between the impulse responses of the source-filter sub-models in the time domain (Equation A.1) or their multiplication in the frequency domain (i.e., the z domain) as shown in Equations A.2, A.3 and A.4.

\[
s(n) = p(n) \ast (A \ast g(n) \ast v(n) \ast r(n))
\]  

(A.1)

Taking the z-transform of A.1 gives
A.1 Speech production and the source-filter model  

\[ S(z) = P(z) \times (A.G(z) \times V(z) \times R(z)) \]  \hspace{1cm} (A.2)

Letting

\[ H(z) = A.G(z) \times V(z) \times R(z), \]  \hspace{1cm} (A.3)

gives

\[ S(z) = P(z) \times H(z), \]  \hspace{1cm} (A.4)

where,

\( s(n) \): Output speech signal,
\( p(n) \): Excitation signal, (which is a train of pulses in the case of voiced speech or noise-like in the case of unvoiced speech.)
\( g(n) \): Impulse response of the glottal shaping filter,
\( A \): Gain of the glottal shaping filter,
\( v(n) \): Impulse response of the vocal tract filter,
\( r(n) \): Impulse response of the lip radiation filter.

The \( z \)-domain representations of the glottal shaping filter, vocal tract filter and lip radiation filter [39] are shown in Equations A.5, A.6 and A.7:

\[ G(z) = z^{-M_0} \prod_{k=1}^{M_0} b_k M_0 \prod_{k=1}^{M_0} (1 - b_k z) \]  \hspace{1cm} (A.5)

The glottal shaping filter is described as a stable filter (with poles located at the origin of the \( z \)-domain), but is anti-causal (i.e., it has zeros outside the unit circle). This type of filter is referred to as maximum phase, where the constants \( |b_k| < 1 \) with Radius-of-Conversion (ROC): \( |z| \geq 1 \). \( M_0 \) is the number of zeros outside the unit circle.
The vocal tract filter is typically an all pole filter with its poles inside the unit circle as shown in Equation A.6.

\[ V(z) = \frac{1}{\prod_{k=1}^{N_i}(1 - c_k z^{-1})}, \]  

(A.6)

where \(|c_k| < 1\) and a ROC: \(|z| \geq 1\). \(N_i\) is the number of poles assumed for the vocal tract model.

The lip radiation filter is given as:

\[ R(z) = \prod_{k=1}^{M_i}(1 - a_k z^{-1}) \]  

(A.7)

This filter has a pole at the origin and a zero at \(a_k\) inside the unit circle, with \(|a_k| < 1\) and a ROC: \(|z| \geq 1\). The overall filter model \(H(z)\) can be obtained by substituting A.5, A.6 and A.7 in A.3 as follows:

\[ H(z) = \frac{z^{-M_0} \prod_{k=1}^{M_0} b_k \prod_{k=1}^{M_i}(1 - b_k z) \prod_{k=1}^{M_i}(1 - a_k z^{-1})}{\prod_{k=1}^{N_i}(1 - c_k z^{-1})} \]  

(A.8)

A.2 Cepstral analysis

When applied to speech, cepstral analysis (often referred as homomorphic filtering [38]) can be used to separate out the various aspects of the speech production process. Cepstral analysis is defined as the inverse Fourier transform of the logarithm of the FFT of the signal [40]. A number of different sets of cepstral coefficients exist. The term cepstrum typically refers to the set obtained when the logarithm function is applied to the magnitude of the FFT components only. When it is applied to both the magnitude and phase components of speech, the resulting set is referred to as the complex cepstrum. Fourier analysis of the speech signal is used to convert the convolution between the source and filter components in the time domain into a product of their corresponding representations in the frequency domain. The logarithm operator transforms this product operation into a sum of
A.2 Cepstral analysis

both components. The inverse Fourier transform is then used to bring the separated components back into the time domain (quefrency domain) [40, 130]. The resultant cepstral coefficients characterize the slow and fast varying components of speech. Slow varying components (e.g. pitch) get concentrated in the upper part of the cepstral domain, while the fast varying components of speech (e.g. the vocal tract filter) get concentrated in the lower part of the cepstral domain. There is a direct relation between frequency and time in the quefrency domain. By way of example, if the sampling frequency is 8000 Hz and a cepstral coefficient peak is found, say at sample 80, then this would indicate the presence of a frequency component and its value being \((\frac{8000 \text{ Hz}}{80} = 100 \text{ Hz})\). Four types of cepstral coefficients are discussed next.

A.2.1 Complex Cepstral Coefficients (CCCs)

When the complex cepstrum analysis is applied to speech, the components of the speech production model will be separated [40] as shown in Equations A.9 and A.10.

\[
\text{IFFT}[\ln(S(z))] = \text{IFFT}[\ln(P(z) + \ln(H(z)))]
\]

which can be written as

\[
\hat{s}(n) = \hat{p}(n) + \hat{h}(n)
\]

Applying the complex logarithm to \(S(z)\) and then inverse Fourier transforming (Equation (A.9)) produces a set of complex cepstral coefficient, \(\hat{s}(n)\). This is the summation of two subsets of coefficients, \(\hat{h}(n)\) (arising from \(H(z)\)) and \(\hat{p}(n)\) (arising from \(P(z)\)) (Equation (A.10)). The \(\hat{h}(n)\) subset of coefficients is derived through the following steps (Equations A.11 to A.15):
\[
\ln\{H(z)\} = \ln(A) + \ln(z^{-M_o}) + \sum_{k=1}^{M_o} \ln(b_k^{-1}) + \sum_{k=1}^{M_o} \ln(1-b_kz) + \\
\sum_{k=1}^{M_i} \ln(1-a_kz^{-1}) - \sum_{k=1}^{N_i} \ln(1-c_kz^{-1}) \quad (A.11)
\]

Using power series expansion of Equation A.12:

\[
\ln(1-z) = -\sum_{n=1}^{\infty} \frac{z^n}{n}, \text{ for } |z| \leq 1, \quad (A.12)
\]

each term in Equation A.11 can be rewritten as in Equation A.13:

\[
\ln\{H(z)\} = \mathcal{H}(z) = \left[ \ln(A) + \sum_{k=1}^{M_o} \ln(b_k^{-1}) \right] z^0 + \sum_{n=1}^{\infty} \left[ \sum_{k=1}^{N_i} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n} \right] z^{-n} - \sum_{n=1}^{\infty} \left[ \sum_{k=1}^{M_o} \frac{b_k^n}{n} \right] z^n \quad (A.13)
\]

Now by definition,

\[
\mathcal{H}(z) = \sum_{n=-\infty}^{\infty} \hat{h}(n)z^{-n} \quad (A.14)
\]

Comparing the different power of \(z^{-1}\) in (A.13) and (A.14) gives

\[
\hat{h}(n) = \begin{cases} 
\ln(A) + \sum_{k=1}^{M_o} \ln(b_k^{-1}) & , n = 0 \\
\sum_{k=1}^{N_i} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n} & , n > 0 \\
\sum_{k=1}^{M_o} \frac{b_k^n}{n} & , n < 0 
\end{cases} \quad (A.15)
\]

\(\hat{h}(n)\) contain both causal and anti-causal components, the causal components arising from \(V(z)\) and \(R(z)\), specifically their poles and zeros inside the unit circle (i.e., constants \(a_k\) and \(c_k\)). The anti-causal component arise from zeros outside the unit circle of the glottal shaping filter \(G(z)\). Both the causal and anti-causal components decay by a factor of \(\frac{1}{n}\). Though \(\hat{h}(n)\) are referred to as complex cepstral coefficients, they are in fact real.
The other subset of coefficients $\hat{p}(n)$, shown in Equation A.10, can also be derived using the same procedure. The excitation signal of voiced speech ($p(n)$) can be defined as a train of impulses. $p(n)$ and its z transform are shown in Equations A.16 and A.17, respectively.

$$p(n) = \sum_{k=0}^{M} \alpha_k \delta(n - kN_p) \quad (A.16)$$

Taking the z-transform of A.16 gives

$$P(z) = \sum_{k=0}^{M} \alpha_k z^{-kN_p} , \quad (A.17)$$

where, $\alpha_k$ is the amplitude of impulse samples and $N_p$ is the period between these impulses (i.e., pitch period). $P(z)$ can also be written as:

$$P(z) = \prod_{k=1}^{M_i} (1 - d_k z^{-N_p}) \prod_{k=1}^{M_o} (1 - g_k^{-1} z^{-N_p}), \quad (A.18)$$

where, $\delta(n)$ are the impulses, $|d_k|, |g_k| < 1$ and ROC: $|z| \geq 1$. Applying the logarithm operator to $P(z)$ and then using Equation A.12 gives

$$\hat{P}(z) = \ln\{P(z)\} = \sum_{k=1}^{M_i} \ln(1 - d_k z^{-N_p}) + \sum_{k=1}^{M_o} \ln(1 - g_k^{-1} z^{-N_p}) \quad (A.19)$$

Similarly, $\hat{p}(n)$ can be obtained by comparing Equation A.19 to its z transform $\hat{P}(z)$, which results in the following relationships:

$$\hat{p}(n) = \begin{cases} \sum_{k=1}^{M_o} \ln(g_k^{-1}) & , \quad n = 0 \\ -\sum_{k=1}^{M_i} \frac{d_k}{n} & , \quad n = N_p, 2N_p, 3N_p, \ldots \\ \sum_{k=1}^{M_o} \frac{g_k^{-n}}{n} & , \quad n = -N_p, -2N_p, -3N_p, \ldots \end{cases} \quad (A.20)$$

$\hat{p}(n)$ are real and contain both causal and anti-causal components with a decaying factor of $1/n$ and they are repeating at integer multiples of $N_p$. Overall, though, $\hat{h}(n)$ dominates the lower part of $\hat{s}(n)$, whereas $\hat{p}(n)$ dominates the higher part.
A.2.2 Real Cepstral Coefficients (RCCs)

RCC analysis follows the same procedure as that of the complex cepstral coefficients, but only uses the magnitude of speech before applying the logarithm operator. As the procedure of RCCs extraction ignores phase information, it cannot be used to retrieve the original speech. RCCs are obtained by taking the even component of the complex cepstral coefficients \[40\], namely

\[
RCC(n) = \frac{(\hat{s}(n) + \hat{s}(-n))}{2} \quad (A.21)
\]

RCCs are a summation of two subsets of coefficients \(\hat{h}_{RCC}(n)\) and \(\hat{p}_{RCC}(n)\) (Equations A.22 and A.23), \(\hat{h}_{RCC}(n)\) (arising from \(H(z)\)) and \(\hat{p}_{RCC}(n)\) (arising from \(P(z)\)).

Applying A.21 to the complex cepstral \(\hat{h}(n)\) in (A.15) gives

\[
\hat{h}_{RCC}(n) = \begin{cases} 
\ln|A| + \sum_{k=1}^{M} \ln|b_k^{-1}|, & n = 0 \\
\frac{1}{2} \left[ \sum_{k=1}^{N_i} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{d_k^n}{n} - \sum_{k=1}^{M_o} \frac{b_k^{-n}}{n} \right], & n > 0 
\end{cases} \quad (A.22)
\]

and applying it to \(\hat{p}(n)\) in (A.20) also gives

\[
\hat{p}_{RCC}(n) = \begin{cases} 
\sum_{k=1}^{M_o} \ln|g_k^{-1}|, & n = 0 \\
\frac{1}{2} \left[ -\sum_{k=1}^{M_i} \frac{d_k^n}{n} - \sum_{k=1}^{M_o} \frac{g_k^{-n}}{n} \right], & n = N_p, 2N_p, 3N_p, \ldots 
\end{cases} \quad (A.23)
\]

Only causal components of the speech signal are captured by RCCs. Again the pitch information is observed at integer multiples of \(N_p\).

A.2.3 Linear Prediction Cepstral Coefficients (LPCCs)

LPCC analysis is performed on the basis of a simplified form of the source-filter model, which lumps together the glottal shaping, lip radiation and vocal tract filter in an all pole
filter model. With LPCCs, LPC analysis is first performed to model the spectral envelop of speech with an all-pole filter [39] as given in Equation A.24. This is followed by an Inverse Discrete Fourier Transform (IDFT) of the logarithm of $H_{LPC}(z)$. Alternatively, the LPC coefficients can be directly converted to LPCCs using a simple recursive algorithm (see [131] for further details). LPCCs are not used, though, to separate out components of the speech production model.

$$H_{LPC}(z) = \frac{G}{\prod_{k=1}^{p} (1 - c_k z^{-1})}, \quad |c_k| < 1 \tag{A.24}$$

Taking the natural logarithm of A.24 gives

$$h_{LPC}(n) = \begin{cases} \ln(G), & n = 0 \\ \sum_{k=1}^{p} \frac{c_k^n}{n}, & n > 0 \\ 0, & n < 0 \end{cases} \tag{A.25}$$

where $p$ is the number of poles assumed for the LPC filter model. $G$ is the $H_{LPC}(z)$ filter gain. LPCCs arise from the causal components of speech, because they are derived from an all-pole filter, with poles inside the unit circle. Since the LPC analysis lumps together all filters in an all-pole model, they are not expected to perform as well as other types of cepstral coefficients.

### A.2.4 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC analysis focuses on the perceptually relevant aspects of the speech spectrum. The speech signal is converted into the frequency domain using the Discrete Fourier Transform (DFT). The next step is estimating how much energy exists in various regions of the frequency domain. This is motivated by the fact that the human ear responds non-linearly at different frequencies. This non-linear response to frequencies is best represented by the Mel-scale as shown in Equation (A.26).

$$M(f) = 1125 \ln(1 + \frac{f}{700}), \tag{A.26}$$
where \( f \) is the frequency (Hz) and \( M(f) \) is the equivalent Mel-scale frequency. The energy is then estimated over a set of overlapped Mel-filter banks by computing the power spectrum of the speech signal and then summing up the energies in each filter bank region. Once the filter bank energies are computed, the logarithm operator is applied. Unlike CCCs, which use IFFT in their last step, the MFCC extraction performs the Discrete Cosine Transform (DCT) on the logarithm of the energies computed. The resultant set \( \text{MFCC}(n) \) is causal as given in Equation (A.27)

\[
\text{MFCC}(n) = \frac{1}{R} \sum_{r=1}^{R} \ln[MF(r)]\cos\left[\frac{2\pi}{R}(r + \frac{1}{2})n\right],
\]

(A.27)

where, \( \text{MFCC}(n) \) is the \( n \)th MFCC coefficient extracted from a particular speech segment using \( R \) filter banks \( (MF) \) [39]. It is a common practice to use deltas and delta-deltas along with MFCCs to capture the dynamic aspects of the speech signal [126]. These are simply the first and second order derivatives of the MFCCs over a range of short-term speech frames. Given that the CCCs arise from taking the complex logarithm of the FFT of \( s(n) \) (both magnitude and phase), whereas MFCCs arise from taking the logarithm of the magnitude of the FFT only, a set of CCCs should potentially contain more speaker-specific information than MFCCs and thus their discriminatory power is expected to be higher.

A.3 Formants and Formant Trajectories (FTs)

Speech formants are identified by locating the dominant peaks of the spectral envelope [42]. The changes in formant values from one speech segment to another are the FTs. FTs capture the dynamic aspect of speech and they are used in a wide range of speech applications [132, 133]. Though speech formants are typically manually extracted, the extraction process has been automated in this study and was found to produce accurate results, when compared to Praat [134]. Firstly, the speech signal is segmented into frames of 20ms and then a tenth order LPC analysis is performed on each frame. The denominator roots of the LPC filter model are then calculated. The roots of each conjugate pair in the
denominator corresponds to one of the speech formants. The angle ($\theta_i$) of each complex root is then computed and the $ith$ formant ($f_i$) is estimated using the following formula:

$$f_i = \frac{\theta_i}{2\pi} f_s,$$

where $f_s$ is the sampling frequency. Each FT was then modelled using an $m$ order Discrete Cosine Transform (DCT) function, where the zeroth DCT coefficient represents the mean value of a FT.
Appendix B

The CDMA-Codec Platform

B.1 Platform usage for simulating mobile phone speech

This software platform was developed as an alternative approach for simulating mobile phone speech, in contrast to the more conventional way of transmitting speech across an actual network. Key aspects of these networks which directly impact the speech signal have been previously discussed in Chapter 5. This platform makes use of the publicly available software simulations of the EVRC-B codec [135]. The layout of this platform is shown in Figure B.1. It incorporates functionality which can be used to operate the speech codec under a wide range of possible scenarios in the CDMA networks. It also has the capability to process, view or listen to multiple or single speech files (.wav format). Using this platform, any existing speech database can be transformed into CDMA-quality speech by adjusting the platform settings in accordance with a desired scenario. The input speech samples must be sampled at 8 kHz and 16 bit digitized.

As shown in Figure B.1, this platform comprises five different panels. These are: (i) Codec Settings, (ii) ADR Mode Files Generator, (iii) Background Noise Adder, (iv) Encode/Decode, and (v) Wav File Player. The Codec Settings panel controls the general behaviour of the EVRC, where the settings can be adjusted in accordance with different scenarios in the CDMA network. Though this platform allows access to various features (or settings) of EVRC, many of those have not been changed from their default values when
simulating mobile phone speech. This is because in an actual mobile phone call, such settings are not changed by the CDMA network, but rather they are used by the developers to test the performance of their codec. The key settings that have been used for coding the speech files in this research are: (i) Anchor Operating Point (AOP), (ii) FL percentage, and (iii) ADR entry for every frame (i.e., switching between AOPs).

The ADR Mode Files Generator has been used for simulating the impact of DRC. Each ADR mode file generated contains a combination of bit rates which are randomly selected in accordance with certain channel capacity conditions thereby resulting in different qualities of speech, as shown in Figure B.2 (details of this will be discussed next).

The Background Noise Adder panel has been used to add different types of BN at different SNR levels to the target speech files, prior to processing them with EVRC. Upon adjusting the codec settings, the Encode/Decode panel was then used to encode and decode all the target speech files. Finally, the Wav File Player panel was used to listen to and compare the original and coded wav files. The start and end times can be changed when listening to different parts of the speech signal. However, this panel can only be used when processing a single wav file. Further details about simulating the impact of DRC, FL and BN at the transmitting end are discussed next.

### B.1.1 Simulating the impact of DRC

Before discussing the specifics of this aspect, it is important to briefly overview the process of DRC. The EVRC can be instructed by the CDMA network to achieve a certain ADR every 20 ms and this results in switching between the three anchor operating points (OP0, OP1 and OP2). As a result of this process, the resultant speech quality can be broadly classified into one of three categories: low-, medium- and high-quality speech. Switching between OP0 and OP1 produces high-quality speech and switching between OP1 and OP2 produces relatively low-quality speech. Under medium-capacity conditions (i.e., the cellular site is neither congested nor has few users), medium-quality speech is produced, which results in switching between all three anchor operating points.\(^1\)

\(^1\)The mapping of ADR values into anchor operating points has been previously shown in Figure 5.2.
Figure B.1: The EVRC codec platform layout and settings.
As shown in Figure B.2, for each speech quality, the ADR values are sampled at different points of the ADR continuous range (refer to Figure 5.2) according to the following: (i) lower-half (low-quality speech), (ii) entire range (medium-quality speech), and (iii) upper-half (high-quality speech). The ADR values chosen for each of these ranges were as follows [2]:

- Low-quality speech: (4.8, 5.8, 6.2, 6.6) kbps.
- Medium-quality speech: (4.8, 5.8, 6.2, 6.6, 7.0, 7.5, 8.5, 9.6) kbps.
- High-quality speech: (7.0, 7.5, 8.5, 9.6) kbps.

Figure B.2: Block diagram illustrating how speech files have been processed corresponding to different capacity conditions.

The selection of discrete values ensures a better coverage of these ranges and thus enforces switching between the AOPs in order to reflect more realistic scenarios in the network. We know for instance that ADR values for medium speech quality can take on any value between 4.8 and 9.6 kbps [111], but random selection from such a continuous range can result in these being more concentrated at one part of the range than another. Using discrete values increases the likelihood of switching between anchor points and ensures a better coverage of the range.
This platform incorporates a facility to create ADR files in accordance with the above speech qualities (or capacity conditions). The process of simulating DRC in this study was as follows. First step, the total number of ADR mode files was specified. This information is used by the platform to create a pool of ADR mode files that corresponds to a particular speech quality, which must be selected from the speech quality menu as shown in Figure B.1. When the coding process for multiple speech files starts, using the Encode/Decode panel, an ADR mode file will then be randomly selected for each speech file from the generated pool of ADR files. As a result, different speech files will be coded using different ADR mode files, which reflects more realistic scenarios in the network than if, for example, those were coded using the same ADR mode file. It needs to be noted here that this facility allows studying the impact of DRC on speech in isolation from FL and BN at the transmitting end.

### B.1.2 Simulating the impact of FL

The original software implementations of the EVRC have had to be modified in order to allow introducing lost frames in the path between the encoder and decoder. In terms of the modifications that have had to be made to the original software codec simulations in order to implement this feature, the frame loss mechanism was internally activated in the EVRC codec by changing the “data_packet.PACKET_RATE” flag value to “0xE” [111, 2].

For simulating the impact of FL, a percentage of FL was first specified in the Codec Settings. The total number of lost frames is made equal to this percentage multiplied by the number of 20 ms frames found in the target speech file. As mentioned earlier in Chapter 8, FL cannot be studied in isolation from DRC, but rather the functionality of DRC can be constrained to one of the AOPs. The use of fixed AOPs (OP0, OP1 or OP2) prevents switching between different coding algorithms. Therefore, when FL was simulated, the AOP menu in the Codec Settings panel was set to OP0, OP1 or OP2 and the ADR entry for every speech frame was disabled. When the coding process (i.e., encoding and decoding) starts, the temporal locations of lost frames for each speech file is randomly selected by the platform using a uniform random distribution. The choice of this distribution is arbitrary at
this stage, but it was considered a reasonable choice in the absence of information regarding more appropriate distributions for FL. Though, this facility allows for any percentage of FL to be specified, one should, in fact, not exceed 10-15%. This is because, in reality, the CDMA network drops a call if the FL percentage exceeds this range [121].

B.1.3 Simulating the impact of BN

This platform has also been used to add different kinds of BN at various SNR levels prior to processing the speech files, which simulates the presence of BN at the transmitting end in a mobile phone call. With this platform, the impact of BN on speech can be investigated in isolation from FL, but cannot be studied in isolation from DRC. Again, when studying the impact of BN, DRC was restricted to OP0, OP1 or OP2 to ensure that the same coding algorithm was used during the coding process.

The most common types of BN in mobile phone communications are babble, car and street noise, and these noise types have therefore been included in this platform. The BN files used have been acquired from the Soundjay database [124]. However, if the user prefers to use a different set of BN files, they can be simply replaced. It is known that in mobile phone communications, typical SNR levels at the transmitting end vary from 9 to 20 dB [136, 2, 111]. Though the process of NS is enabled by default in the EVRC, this platform allows enabling or disabling this feature. This can be quite useful in order to understand whether the distortion of speech has been caused by the coding process itself or is as a result of NS.