

**The Effect of Acoustic  
Variability  
on  
Automatic  
Speaker Recognition  
Systems**

**John Nash**

**Submitted in fulfillment of the requirements for  
the degree of Doctor of Philosophy**

**The University of York  
Language and Linguistic Science**

**Submitted September 2019**





# Abstract

---

This thesis examines the influence of acoustic variability on automatic speaker recognition systems (ASRs) with three aims.

- i. To measure ASR performance under 5 commonly encountered acoustic conditions;
- ii. To contribute towards ASR system development with the provision of new research data;
- iii. To assess ASR suitability for forensic speaker comparison (FSC) application and investigative/pre-forensic use.

The thesis begins with a literature review and explanation of relevant technical terms. Five categories of research experiments then examine ASR performance, reflective of conditions influencing speech quantity (inhibitors) and speech quality (contaminants), acknowledging quality often influences quantity. Experiments pertain to: net speech duration, signal to noise ratio (SNR), reverberation, frequency bandwidth and transcoding (codecs). The ASR system is placed under scrutiny with examination of settings and optimum conditions (e.g. matched/unmatched test audio and speaker models). Output is examined in relation to baseline performance and metrics assist in informing if ASRs should be applied to suboptimal audio recordings.

Results indicate that modern ASRs are relatively resilient to low and moderate levels of the acoustic contaminants and inhibitors examined, whilst remaining sensitive to higher levels. The thesis provides discussion on issues such as the complexity and fragility of the speech signal path, speaker variability, difficulty in measuring conditions and mitigation (thresholds and settings). The application of ASRs to casework is discussed with recommendations, acknowledging the different modes of operation (e.g. investigative usage) and current UK limitations regarding presenting ASR output as evidence in criminal trials.

In summary, and in the context of acoustic variability, the thesis recommends that ASRs could be applied to pre-forensic cases, accepting extraneous issues endure which require governance such as validation of method (ASR standardisation) and population data selection. However, ASRs remain unsuitable for broad forensic application with many acoustic conditions causing irrecoverable speech data loss contributing to high error rates.

# Contents

---

<b>Abstract</b>	<b>3</b>
<b>Contents</b>	<b>4</b>
<b>Tables, Figures and Illustrations</b>	<b>8</b>
<b>Accompanying Material</b>	<b>12</b>
<b>Acknowledgements</b>	<b>13</b>
<b>Declaration</b>	<b>14</b>
<b>Chapter 1 Introduction</b>	<b>15</b>
1.1 <i>Speaker Recognition</i>	15
1.2 <i>Research Aims</i>	16
1.3 <i>Thesis outline</i>	17
<b>Chapter 2 Literature Review</b>	<b>19</b>
2.1 <i>Speaker Comparison Methodologies</i>	19
2.2 <i>Forensic Speaker Comparison</i>	21
2.3 <i>Automatic Speaker Recognition Systems</i>	23
<b>Chapter 3 From Speaker Source to Analytical Destination</b>	<b>30</b>
3.1 <i>Speech Production</i>	30
3.1.1 <i>Voice Quality</i>	34
3.1.2 <i>Sound Pressure Levels</i>	37
3.2 <i>Intrinsic and Extrinsic Variability</i>	38
3.3 <i>Audio Recording and The Signal Path</i>	41
3.3.1 <i>Audio Capture</i>	41
3.3.2 <i>Digital Recording and Sampling</i>	42
3.3.3 <i>Transcoding (Codec)</i>	45
3.3.4 <i>Bit Rate</i>	45
3.3.5 <i>Spectrogram Analysis</i>	45
3.3.6 <i>Signal to Noise Ratio</i>	47
3.3.7 <i>Reverberation</i>	47
3.3.8 <i>Frequency Bandwidth</i>	48
3.4 <i>Automatic Speaker Recognition Systems</i>	49
3.4.1 <i>Speech Detection</i>	49
3.4.2 <i>Diarisation/Speaker Separation</i>	50
3.4.3 <i>Feature Extraction and Mel Frequency Cepstral Coefficients</i>	52
3.4.4 <i>Long Term Formant Distribution</i>	55
3.4.5 <i>Statistical Modelling</i>	56
3.4.5.1 <i>Gaussian Mixture Models</i>	56
3.4.5.2 <i>i-Vectors and Statistical Modelling</i>	58
3.4.6 <i>Speaker Model or Voiceprint?</i>	59
3.4.7 <i>Normative Data/Background Population</i>	60
3.5 <i>Automatic Speaker Recognition Output and Performance Measurement</i>	61
3.5.1 <i>Likelihood Ratio and Bayes' Theorem</i>	61
3.5.2 <i>Verbal equivalence scales</i>	64
3.5.3 <i>Likelihood Ratio and Log Likelihood Ratio Plots</i>	66
3.5.4 <i>System Accuracy and Precision</i>	67
3.5.5 <i>Cost of Log Likelihood Ratio</i>	68
3.5.6 <i>Zoo plots</i>	69
3.5.6.1 <i>Zoo Plots and Inter/Intra Variability</i>	72
3.5.6.2 <i>Performance Measurements (False Accept Rate and False Reject Rate)</i>	74
3.6 <i>Automatic Speaker Recognition Use Case Examples</i>	74
3.7 <i>Summary</i>	76
<b>Chapter 4 Research Questions</b>	<b>78</b>
<b>Chapter 5 Equipment and Recordings</b>	<b>81</b>
5.1 <i>Software</i>	81

5.1.1	<i>Audio Applications and Scripts</i> .....	81
5.1.2	<i>Vocalise and iVocalise Software</i> .....	82
5.1.3	<i>Bio-Metrics Software</i> .....	83
5.2	<i>Summary of Hardware</i> .....	83
5.3	<i>Speech Corpora</i> .....	84
<b>Chapter 6</b>	<b>Preliminary Testing</b> .....	<b>85</b>
6.1	<i>Objectives</i> .....	85
6.2	<i>Questions</i> .....	85
6.3	<i>Data Preparation and Materials</i> .....	86
6.3.1	<i>User Mode</i> .....	87
6.4	<i>Selection of Normative Data</i> .....	88
6.4.1	<i>Additional Data</i> .....	88
6.5	<i>Preliminary Test Results</i> .....	89
6.5.1	<i>Automatic Speaker Recognition Settings and Equal Error Rate Results</i> .....	89
6.5.2	<i>Cepstral and Formant System Comparison</i> .....	90
6.5.3	<i>Zoo Plot Analysis</i> .....	91
6.5.4	<i>Voice Quality and Accent Data</i> .....	92
6.5.5	<i>Additional Analysis (Speaker 012)</i> .....	95
6.5.6	<i>Signal to Noise Ratio Test Results</i> .....	96
6.5.7	<i>Net Duration Test Results</i> .....	97
6.5.8	<i>Zoo Plot Position</i> .....	100
6.6	<i>Responses to Questions</i> .....	100
6.7	<i>Conclusion</i> .....	101
<b>Chapter 7</b>	<b>Net Duration</b> .....	<b>103</b>
7.1	<i>Introduction</i> .....	103
7.2	<i>Background</i> .....	104
7.3	<i>Additional Definition of Terms</i> .....	105
7.4	<i>Literature Review</i> .....	106
7.5	<i>Questions and Hypotheses</i> .....	112
7.6	<i>Methodology</i> .....	113
7.7	<i>Results</i> .....	114
7.8	<i>Responses to Research Questions</i> .....	120
7.9	<i>Discussion and Practical Recommendations</i> .....	126
<b>Chapter 8</b>	<b>Signal to Noise Ratio</b> .....	<b>129</b>
8.1	<i>Background</i> .....	129
8.2	<i>Literature Review</i> .....	130
8.2.1	<i>Vocal Effort and Signal to Noise Ratio</i> .....	135
8.2.2	<i>Signal to Noise Ratio Estimation</i> .....	136
8.3	<i>Questions and Hypotheses</i> .....	137
8.4	<i>Methodology</i> .....	138
8.5	<i>Results</i> .....	140
8.6	<i>Findings</i> .....	148
8.7	<i>Discussion</i> .....	150
<b>Chapter 9</b>	<b>Reverberation</b> .....	<b>152</b>
9.1	<i>Introduction</i> .....	152
9.2	<i>Literature Review</i> .....	156
9.3	<i>Questions and Hypotheses</i> .....	161
9.3.1	<i>Additional Experiment</i> .....	162
9.4	<i>Methodology</i> .....	163
9.4.1	<i>Reverberation Modeling</i> .....	163
9.4.2	<i>Data Preparation</i> .....	164
9.4.3	<i>Normative Data, Gaussian Mixture Model System</i> .....	165
9.4.4	<i>Normative Data, i-vector System</i> .....	165
9.5	<i>Results</i> .....	166
9.5.1	<i>Observations</i> .....	170
9.5.2	<i>System Accuracy Results</i> .....	176
9.5.3	<i>Results from Normative Sessions 2 and 3</i> .....	178

9.5.4	<i>Speech Detection Results</i> .....	181
9.6	<i>Discussion of Results</i> .....	182
9.6.1	<i>Responses to Questions</i> .....	182
9.6.2	<i>Voice Activity Detection</i> .....	184
9.7	<i>Recommendations</i> .....	184
9.8	<i>Discussion and Future Research</i> .....	186
<b>Chapter 10</b>	<b>Frequency Bandwidth</b> .....	<b>187</b>
10.1	<i>Introduction</i> .....	187
10.2	<i>Context</i> .....	187
10.2.1	<i>Background</i> .....	187
10.2.2	<i>Literature Review</i> .....	188
10.3	<i>Questions and Hypotheses</i> .....	192
10.4	<i>Methodology</i> .....	193
10.4.1	<i>Baseline Corpus</i> .....	193
10.4.2	<i>Automatic Speaker Recognition Systems and Additional Materials</i> .....	193
10.4.3	<i>Test Audio and Speaker Models</i> .....	194
10.4.4	<i>Normative Data</i> .....	195
10.4.5	<i>Automatic Speaker Recognition Systems</i> .....	195
10.4.6	<i>Data and List of Experiments</i> .....	196
10.5	<i>Results</i> .....	196
10.5.1	<i>iVocalise, i-vector System Results</i> .....	197
10.5.2	<i>Vocalise, Gaussian Mixture Model System Results</i> .....	198
10.5.3	<i>Zoo Plots</i> .....	200
10.6	<i>Responses to Questions</i> .....	207
10.6.1	<i>Summary of Results</i> .....	210
10.7	<i>Discussion and Practical Application</i> .....	210
<b>Chapter 11</b>	<b>Transcoding</b> .....	<b>213</b>
11.1	<i>Introduction</i> .....	213
11.2	<i>Background</i> .....	213
11.3	<i>Literature Review</i> .....	214
11.4	<i>Questions and Hypotheses</i> .....	220
11.5	<i>Methodology and Materials</i> .....	221
11.5.1	<i>List of Codecs for Comparison</i> .....	221
11.5.2	<i>Automatic Speaker Recognition Systems</i> .....	223
11.6	<i>Results</i> .....	224
11.7	<i>Responses to Questions</i> .....	243
11.8	<i>Findings</i> .....	244
11.9	<i>Additional Tone Experiment</i> .....	246
11.10	<i>Discussion</i> .....	250
<b>Chapter 12</b>	<b>Discussion</b> .....	<b>253</b>
12.1	<i>Summary of Automatic Speaker Recognition Performance</i> .....	253
12.1.1	<i>Tipping Points and Acceptability Criteria</i> .....	253
12.1.2	<i>Automatic Speaker Recognition Performance Metrics</i> .....	254
12.1.3	<i>Opportunities for Automatic Speaker Recognition Improvements</i> .....	255
12.1.4	<i>Technical Quality Assessments</i> .....	255
12.1.5	<i>Mismatched Conditions</i> .....	256
12.1.6	<i>Population/Normative Data</i> .....	257
12.1.7	<i>Operator Training and Standards</i> .....	258
12.2	<i>Practical Recommendations</i> .....	259
12.2.1	<i>Net Duration Recommendations</i> .....	259
12.2.2	<i>Signal to Noise Ratio Recommendations</i> .....	262
12.2.3	<i>Reverberation Recommendations</i> .....	263
12.2.4	<i>Frequency Bandwidth Recommendations</i> .....	264
12.2.5	<i>Transcoding/Codec Recommendations</i> .....	264
12.2.6	<i>Control Corpora and Test Data</i> .....	265
12.2.7	<i>Automatic Speaker Recognition System Recommendations</i> .....	265
12.2.8	<i>Performance Metrics</i> .....	266

12.2.9	<i>Auditory Analysis</i> .....	266
12.3	<i>Should Automatic Speaker Recognition Transition to Forensic Use?</i> .....	266
<b>Chapter 13</b>	<b>Future Research</b> .....	<b>269</b>
13.1	<i>Combining Acoustic Conditions</i> .....	269
13.2	<i>Modeling Automatic Speaker Recognition System Environments</i> .....	269
13.2.1	<i>Applying Big Data for Mismatch Compensation</i> .....	269
13.3	<i>Pre-Analysis Audio Enhancement</i> .....	270
13.4	<i>Feature Extraction Methods and System Fusion</i> .....	270
13.4.1	<i>Automatic Speech Recognition</i> .....	272
13.5	<i>Automated Audio Quality Measurement</i> .....	272
13.6	<i>Alternate Approaches to Speaker Model Generation</i> .....	272
13.7	<i>X-Vector Automatic Speaker Recognition Systems</i> .....	273
<b>Chapter 14</b>	<b>Conclusion</b> .....	<b>274</b>
<b>Appendices</b>	.....	<b>277</b>
<b>Appendix A</b>	.....	<b>278</b>
<b>Appendix B</b>	.....	<b>280</b>
<b>Appendix C</b>	.....	<b>284</b>
<b>Appendix D</b>	.....	<b>304</b>
<b>Appendix E</b>	.....	<b>313</b>
<b>Appendix F</b>	.....	<b>333</b>
<b>Appendix G</b>	.....	<b>334</b>
<b>Appendix H</b>	.....	<b>335</b>
<b>Appendix I</b>	.....	<b>338</b>
<b>Appendix J</b>	.....	<b>340</b>
<b>Appendix K</b>	.....	<b>342</b>
<b>Appendix L</b>	.....	<b>343</b>
<b>Abbreviations</b>	.....	<b>344</b>
<b>Bibliography</b>	.....	<b>347</b>

# Tables, Figures and Illustrations

---

<b>Figure 2.1:</b> Morrison et al. (2016: p.94). Summary of LEA speaker comparison methodologies	20
<b>Figure 2.2:</b> Shaver and Acken (2016) early speaker comparison timeline (reproduced)	26
<b>Figure 2.3:</b> Summary of ideal speaker comparison conditions	27
<b>Figure 2.4:</b> ASR process, reproduced and adapted from Campbell (1997: p.1438)	28
<b>Figure 2.5:</b> Key conclusion points reproduced from Becker et al. (2012: pp.5-6)	29
<b>Figure 3.1:</b> Construction of the larynx. Anterior and lateral views	31
<b>Figure 3.2:</b> Cross section of the vocal folds (and glottis)	32
<b>Figure 3.3:</b> Detailed sagittal section of the respiratory tract	33
<b>Table 3.4:</b> Pisoni and Remez (2004: p.347) early evolution of VQ description	34
<b>Table 3.5:</b> San Segundo and Mompean (2017: p.644[e]23) VPA Table	35
<b>Figure 3.6:</b> Fant (1959: p.4) Sound pressure level and vowel and consonant frequency	38
<b>Table 3.7:</b> Examples of intrinsic speaker variability	39
<b>Table 3.8:</b> Examples of extrinsic speaker variability	40
<b>Figure 3.9:</b> Analogue to digital conversion and sample rate	43
<b>Figure 3.10:</b> Analogue to digital conversion and bit depth	44
<b>Figure 3.11:</b> iZotope RX Advanced spectrogram and waveform views	46
<b>Figure 3.12:</b> Visual representations of speech in Praat, amplitude and spectrogram	46
<b>Figure 3.13:</b> An example of modern, standalone, diarisation software (Clever by OWR)	51
<b>Figure 3.14:</b> Mel and Hz scales	52
<b>Figure 3.15:</b> Summary of the MFCC feature extraction process	53
<b>Figure 3.16:</b> Conceptual example of MFCC values/heatmap from Lode et al. (2018: p.5)	54
<b>Figure 3.17:</b> Illustration of 5 Gaussian components forming a GMM (Dulal, 2014)	57
<b>Figure 3.18:</b> Bayes' theorem from Drygajlo et al./ENFSI (2015)	62
<b>Figure 3.19:</b> Typicality, similarity and calculation of LR, from Morrison (2009)	63
<b>Table 3.20:</b> Verbal equivalence scale from Rose (2002: p.61)	64
<b>Table 3.21:</b> ENFSI Verbal equivalence scale (ENFSI 2015: p17)	65
<b>Figure 3.22:</b> LR Plot example from ENFSI standards (2015: p.19) + additional annotation	66
<b>Figure 3.23:</b> Accuracy and Precision explanation	68
<b>Figure 3.24:</b> Dunstone and Yager (2009) and Doddington's classification systems	70
<b>Figure 3.25:</b> Example zoo plot, showing categories	71
<b>Figure 3.26:</b> Example of an OWR Bio-Metrics zoo plot with fat and thin animals	73
<b>Table 3.27:</b> Typical examples of ASR use cases	75
<b>Figure 3.28:</b> Speaker comparison timeline, evolution from 2005 to 2019	77
<b>Table 6.1:</b> Summary of preliminary EER% results. Vocalise, DyViS, telephone channel	89
<b>Figure 6.2:</b> Bio-Metrics LR plot. 100 SM x 300 TA, MFCC GMM-UBM	90
<b>Figure 6.3:</b> Bio-Metrics LR plot. 100 SM x 300 TA, LTFD GMM-UBM	91
<b>Table 6.4:</b> Vocalise ASR, MFCC. EER 1.2441%: Zoo plot categories by speaker number	91
<b>Table 6.5:</b> Vocalise ASR, LTFD. EER 6.0219%: Zoo plot categories by speaker number	92
<b>Figure 6.6:</b> Zoo plot 100 SM x 300 TA, GMM-UBM MFCC. VQ Data 1 Example	93
<b>Figure 6.7:</b> Zoo plot 100 SM x 300 TA, GMM-UBM LTFD. VQ Data 2 Example	94
<b>Figure 6.8:</b> Zoo plot 100 SM x 300 TA, GMM-UBM MFCC	95
<b>Figure 6.9:</b> 30,000 comparisons MFCC Vocalise. Blue line shows TP scores	95
<b>Figure 6.10:</b> Zoo plot 100 SM x 300 TA, MFCC GMM-UBM. Net duration and SNR tests	98
<b>Figure 6.11:</b> Zoo plot 100 SM x 300 TA, LTFD GMM-UBM. Net duration and SNR tests	99
<b>Table 7.1:</b> Training and truncated test data part I. Kanagasundaram et al. (2011: p.2344)	107
<b>Table 7.2:</b> Training and truncated test data part II. Kanagasundaram et al. (2011: p.2344)	107
<b>Figure 7.3:</b> Poddar, Sahidullah and Saha results 2015 (GMM and i-vector/PLDA)	109
<b>Figure 7.4:</b> Poddar, Sahidullah and Saha 2D i-vectors on low duration files (2018: p.94)	109
<b>Figure 7.5:</b> Ma et al. (2017: p.405) 2D i-vectors and short/long net duration	111

<b>Table 7.6:</b> Net duration experiment 1, GMM-UBM and i-vector/PLDA results, matched .....	114
<b>Figure 7.7:</b> Net duration experiment 1(a and b). EER%, i-vector and GMM-UBM.....	115
<b>Figure 7.8:</b> Net duration experiment 1a and b. Cllr, i-vector and GMM-UBM, matched .....	115
<b>Table 7.9:</b> Net duration experiment 2. EER% Results. iVocalise, i-vector system.....	116
<b>Table 7.10:</b> Net duration experiment 2. Cllr Results. iVocalise, i-vector system .....	117
<b>Figure 7.11:</b> Experiment 3. iVocalise 50 speakers, 1m SM, 2x 1m TA comparisons circled .....	118
<b>Figure 7.12:</b> Experiment 3, iVocalise 50 speakers, 1m SM, 2x 20s TA comparisons circled.....	119
<b>Figure 7.13:</b> iVocalise results at 3s SM x 3s TA (x2). 100 speakers, 4 outlier speakers circled .	122
<b>Figure 7.14:</b> iVocalise results at 60s SM x 60s TA (x2). 100 speakers, 4 outliers as in 7.13.....	123
<b>Figure 7.15:</b> Praat spectrogram view of speakers 020 (best) and 025 (worst) 3s.....	125
<b>Table 8.1:</b> Influence of SNR on GMM ASR system. Togneri and Pullella (2011: p.37) .....	130
<b>Figure 8.2:</b> Results from Athulya, Vinashankar and Sathidevi (2017: p.5).....	132
<b>Figure 8.3:</b> Li and Mak (2016: p. 5566) shift of mean i-vectors with SNR reduction .....	133
<b>Table 8.4:</b> Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) noise settings .....	134
<b>Figure 8.5:</b> Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) noise results .....	135
<b>Figure 8.6:</b> Pink noise (left) and white noise generators, showing spectral tilt .....	139
<b>Table 8.7:</b> SNR Experiments detailing noise types and settings .....	139
<b>Table 8.8:</b> SNR Experiments, iVocalise, i-vector/PLDA, results (next 2 pages).....	140
<b>Figure 8.9:</b> I-vector ASR. White noise matched SM and TA EER% results.....	143
<b>Figure 8.10:</b> I-vector ASR. Pink noise matched SM and TA EER% results .....	143
<b>Figure 8.11:</b> I-vector ASR. White noise non-matched SM and TA EER% results.....	144
<b>Figure 8.12:</b> I-vector ASR. Pink noise non-matched SM and TA EER% results .....	144
<b>Table 8.13:</b> WADA SNR Estimates for 100 x DyViS speakers, task 1 (SM, baseline) .....	145
<b>Figure 8.14:</b> Zoo plot baseline results. Lowest 10% of speakers, WADA SNR.....	146
<b>Figure 8.15:</b> Zoo plot -20db RMS White Noise. Lowest 10% of speakers, WADA SNR .....	146
<b>Figure 8.16:</b> Zoo plot baseline results. Highest 10% of speakers, WADA SNR .....	147
<b>Figure 8.17:</b> Zoo plot -20db RMS White Noise. Highest 10% of speakers, WADA SNR.....	147
<b>Table 8.18:</b> Summary of results from audio enhancement experiments .....	149
<b>Figure 8.19:</b> Praat spectrogram. Speaker 2 SM (1.875s) baseline.....	150
<b>Figure 8.20:</b> Praat spectrogram. Speaker 2 SM (1.875s) -20dbRMS white noise.....	150
<b>Figure 9.1:</b> Baseline data, no reverberation applied. DyViS Speaker 120.....	154
<b>Figure 9.2:</b> Car reverberation applied (RT60 = 0.60). DyViS Speaker 120 .....	154
<b>Figure 9.3:</b> Living Room reverberation applied (RT60 = 0.70). DyViS Speaker 120 .....	155
<b>Figure 9.4:</b> Hall reverberation applied (RT60 = 1.40). DyViS Speaker 120 .....	155
<b>Figure 9.5:</b> Summary of results reproduced from Avila et al. (2015: p.4) .....	159
<b>Table 9.6:</b> Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) Reverberation settings.....	160
<b>Figure 9.7:</b> Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) Reverberation results.....	161
<b>Table 9.8:</b> IR-L Reverberation settings selected for the experiments.....	164
<b>Table 9.9:</b> GMM-UBM Results. Matched conditions .....	166
<b>Table 9.10:</b> GMM-UBM Results. Unmatched conditions .....	167
<b>Table 9.11:</b> I-vector/UBM, TV, LDA+PLDA results. Matched conditions .....	167
<b>Table 9.12:</b> I-vector/UBM, TV, LDA+PLDA results. Unmatched conditions .....	168
<b>Figure 9.13:</b> Influence of reverberation on i-vector/UBM, TV, LDA+PLDA ASR .....	168
<b>Figure 9.14:</b> Influence of reverberation on i-vector/UBM, TV, LDA+PLDA EER% .....	169
<b>Figure 9.15:</b> Influence of reverberation on GMM-UBM ASR EER%.....	169
<b>Figure 9.16:</b> Zoo plot of baseline data (i-vector/UBM, TV, LDA+PLDA1).....	172
<b>Figure 9.17:</b> Zoo plot of Living Room results, matched conditions, i-vector, PLDA 3 .....	173
<b>Figure 9.18:</b> LR Plot, Baseline data, matched conditions, PLDA session 3 .....	174
<b>Figure 9.19:</b> LR Plot, Living Room data, matched conditions, PLDA session 3.....	175
<b>Figure 9.20:</b> LR Plot, Hall data, matched conditions, PLDA session 3 .....	175
<b>Table 9.21:</b> I-vector ASR tests (PLDA session 1). Examination of Cllr .....	177
<b>Table 9.22:</b> Summary results from reverberation experiments PLDA2, matched .....	178
<b>Table 9.23:</b> Summary results from reverberation experiments PLDA2, unmatched.....	179
<b>Table 9.24:</b> Summary results from reverberation experiments PLDA3, matched .....	179
<b>Table 9.25:</b> Summary results from reverberation experiments PLDA3, unmatched.....	180

<b>Table 9.26:</b> EER% Optimal performance across PLDA sessions 1, 2 and 3 .....	180
<b>Table 9.27:</b> VAD Results. Matched conditions, PLDA session 3, VAD Off .....	181
<b>Table 9.28:</b> VAD Results. Unmatched conditions, PLDA 3, VAD Off .....	182
<b>Table 10.1:</b> Results from NB and WB ASR performance. Pradhan and Prasanna (2011) .....	189
<b>Table 10.2:</b> Matched SM and TA. iVocalise ASR, bespoke PLDA Results.....	197
<b>Figure 10.3:</b> Matched SM and TA, i-vector ASR, bespoke PLDA. Mean H0, H1 .....	197
<b>Table 10.4:</b> Unmatched SM and TA, i-vector ASR, bespoke PLDA .....	197
<b>Figure 10.5:</b> Unmatched SM and TA, i-vector ASR, bespoke PLDA H0, H1 SD.....	198
<b>Table 10.6:</b> Matched SM and TA. Vocalise GMM-UBM, bespoke UBM.....	199
<b>Figure 10.7:</b> Matched SM and TA. GMM-UBM bespoke UBM. Mean H0, H1 .....	199
<b>Table 10.8:</b> Unmatched SM and TA. GMM-UBM, bespoke UBM. ....	199
<b>Figure 10.8b:</b> Unmatched SM and TA. GMM-UBM, bespoke UBM. Mean H0, H1.....	199
<b>Figure 10.9:</b> Zoo plot re frequency bandwidth, 0-16kHz, SR32kHz Matched SM and TA. ....	202
<b>Figure 10.10:</b> Zoo plot re frequency bandwidth, 0-03kHz, SR06kHz Matched SM and TA. ....	203
<b>Figure 10.11:</b> Zoo plot iVocalise 0-11kHz Matched SM and TA. Dove speakers highlighted. ..	204
<b>Figure 10.12:</b> Zoo plot iVocalise 0-4kHz Matched. Dove speakers from 0-11kHz test .....	205
<b>Figure 10.13:</b> Zoo plot iVocalise 0-3kHz Matched. Dove speakers from 0-11kHz test .....	206
<b>Table 10.14:</b> Comparison in low bandwidth speech, EER% (i-vector/PLDA ASR).....	209
<b>Table 10.15:</b> Results from preliminary tests, showing influence on EER% re addition of F4.....	212
<b>Table 11.1:</b> Polacký, Pocta and Jarina EER% results (2016a: p.81) .....	214
<b>Table 11.2:</b> Silovsky, Cerva and Zdansky codecs evaluted (2011: p.206).....	216
<b>Table 11.3:</b> Silovsky, Cerva and Zdansky results (2011: p.207) + annotation .....	217
<b>Table 11.4:</b> Janicki and Staroszczyk codec results (2011: p.296), bold=best % correct .....	218
<b>Table 11.5:</b> Codec types used in experiments with settings.....	221
<b>Table 11.6:</b> Transcoding results. OWR iVocalise, i-vector/PLDA ASR.....	225
<b>Table 11.7:</b> Transcoding results. OWR Vocalise ASR, GMM-UBM .....	228
<b>Figure 11.8:</b> MP3 EER% results i-vector/PLDA .....	231
<b>Figure 11.9:</b> Speex EER% results, i-vector/PLDA .....	231
<b>Figure 11.10:</b> ADPCM EER% results i-vector/PLDA.....	232
<b>Figure 11.11:</b> Opus EER% results i-vector/PLDA.....	232
<b>Figure 11.12:</b> G.711 EER% results i-vector/PLDA .....	233
<b>Figure 11.13:</b> Speex EER% results GMM-UBM.....	233
<b>Figure 11.14:</b> MP3 CBR EER% results GMM-UBM.....	234
<b>Figure 11.15:</b> Opus CBR EER% results GMM-UBM .....	234
<b>Figure 11.16(i):</b> G.711 EER% results GMM-UBM.....	235
<b>Figure 11.16(ii):</b> ADPCM EER% results GMM-UBM.....	235
<b>Figure 11.17(i):</b> Baseline zoo plot. GMM-UBM ASR Vocalise (1 of 2) .....	236
<b>Figure 11.17(ii):</b> Baseline zoo plot. i-vector/PLDA ASR iVocalise (2 of 2).....	237
<b>Figure 11.18(i):</b> Zoo plot Speex 1 results GMM-UBM ASR System (1 of 2).....	238
<b>Figure 11.18(ii):</b> Zoo plot, Speex 1 results, i-vector/PLDA ASR System (2 of 2).....	239
<b>Figure 11.19:</b> Zoo plot, Dialogic ADPCM 16kHz, i-vector/PLDA ASR System.....	240
<b>Figure 11.20(i):</b> LR Plot. Vocalise GMM-UBM MP3 CBR, 128kbps (1 of 4).....	241
<b>Figure 11.20(ii):</b> LR Plot. Vocalise GMM-UBM MP3 CBR, 32kbps (2 of 4).....	241
<b>Figure 11.20(iii):</b> LR Plot. Vocalise GMM-UBM MP3 CBR, 16kbps (3 of 4).....	242
<b>Figure 11.20(iv):</b> LR Plot. Vocalise GMM-UBM MP3 CBR, 08kbps (4 of 4).....	242
<b>Figure 11.21:</b> Influence of transcoding on formants. Baseline and transcoded MP3,8kbps,CBR	247
<b>Figure 11.22:</b> Spectrogram view of test tones pre transcoding (baseline).....	248
<b>Figure 11.23:</b> Spectrogram view of test tones post transcoding (MP3, CBR, 8kbps).....	248
<b>Figure 11.24:</b> Spectrogram view of test tones post transcoding (Speex, quality 8) .....	249
<b>Table 11.25:</b> Mean frequencies, 3 additional transcoding experiments, 4 test tones.....	249
<b>Figure 11.26:</b> Spectrogram examples of codec distortion and data loss.....	251
<b>Figure 12.1:</b> Zimmerman trial (2013) transcript summary (next page).....	260
<b>Table 13.1:</b> Fedila, Bengherabi and Amrouche (2018: p.16734, Table 6).....	271



<b>Appendix B: Table 1:</b> Example VQ Data, Stevens and French (2013) + subsequent analysis....	280
<b>Appendix B: Table 2:</b> Example VQ Data, Stevens and French (2013) and analysis .....	281
<b>Appendix H: Figure 1:</b> OWR Vocalise LTFD. VQ analysis. Speakers with VQ lax larynx .....	336
<b>Appendix H: Figure 2:</b> OWR Vocalise MFCC. VQ analysis. Speakers without VQ breathy ....	337
<b>Appendix I: Table 1:</b> Poddar, Sahidullah and Saha Tables (2015). .....	338
<b>Appendix J: Table 1:</b> Mean test tone frequencies (Praat, baseline) .....	340
<b>Appendix J: Table 2:</b> Mean test tone frequencies (Praat, .mp3 CBR 8kbps) .....	340
<b>Appendix K: Figure 1:</b> iZotope RX Spectrogram. DyViS Speaker 012, task 2 .....	342

# Accompanying Material

---

Two electronic files (.gifs) are presented as supporting material, referenced in this thesis.

- i. NASH\_108045162\_GMM-UBMAnimation\_FreqBandwidth\_matched\_LRPlots.gif
- ii. NASH\_108045162\_GMM-UBMAnimation\_FreqBandwidth\_matched\_Zoos.gif

The .gif files play in a standard web browser. They have been checked for compatibility with Chrome, Safari and Firefox. Older graphics cards/hardware may not support playback.

By their nature, .gif files are lower in quality than the images they are drawn from. Higher quality images (contributory frames) are included in the Appendices for reference.

# Acknowledgements

---

Sincere thanks to Professor Peter French for accepting me as a research student back in simpler times (2011). I have very much appreciated your guidance, patient supervision, constructive criticism, support and excellent coffee throughout. Thank you very much to Dr. Carmen Llamas too and for supporting me from day one to submission.

Thank you to Dr. Dominic Watt and Dr. Philip Harrison for your support. Your technical views, encouragement and feedback have been very much appreciated, particularly in the early stages and with the pilot studies. Thanks to Dr. Richard Rhodes, David Van Der Vloed, fellow research students at York and the countless experts at IAFPA and AES who have assisted through discussion, advice, encouragement and contributing their experiences. The quality and quantity of speech technology research, that is collectively produced by the international community, has been so inspiring. Thanks also to the library at York university and their online resources which are nothing short of amazing.

Thanks to Dr. Anil Alexander, Dr. Finnian Kelly and the team at Oxford Wave Research U.K. Thank you for making your software suite available for research purposes and allowing it to be placed under intense scrutiny. I am grateful for your technical support, advice and in answering emails and texts at all hours. Thank you to Alastair and to Matthew for your coding advice too.

To my parents, thank you so much for your enduring support and encouragement throughout the years. Finally, to my very best friend and incredibly tolerant wife for supporting me throughout the whole journey. I will never be able to thank you enough. You sacrificed so much to help me get this far and you are simply amazing.

# Declaration

---

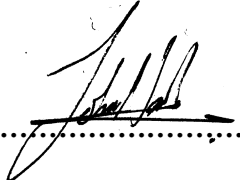
I declare that this thesis is a presentation of original work and I am the sole author. This work has not been presented for an award at this, or any other, University. All sources are acknowledged as references.

During completion of this thesis the following presentation was delivered and co-presented with research colleagues at IAFPA (2014):

Alexander, A., Forth, O., Nash, J. and Yager, N. (2014). *Zoo plots for speaker recognition with tall and fat animals.*

This thesis is less than 80k words excluding the Bibliography, Appendices and Plates/Figures (12k)

**John Nash**

Signature.....

**30<sup>th</sup> of September 2019**  
Date.....

# Chapter 1 Introduction

---

Chapter 1 places the research experiments conducted in this thesis into context. The fundamentals of speaker comparison and automatic speaker recognition systems (ASR<sup>§</sup>s) are introduced. The objectives of the research experiments are stated and the thesis outline is presented. For the scope of this thesis, ASR use focuses on law enforcement/investigative application and the potential use of ASR evidence in criminal trials in the UK.

## 1.1 Speaker Recognition

This section introduces the basic principles of speaker comparison. Further detail is also provided in the literature review and technical terms are explained in Chapter 3.

Attributing speech to speaker is a basic human function of communication. If speech is the only data available for determining identity and visual references are not available (e.g. telephone, audio recording) then several complex processes must take place. For humans, and if we consider naïve listeners rather than experts, this process relies on familiarity and memory. Self-identification from the speaker(s) may also occur within the content of the conversation (assuming that the speaker is being truthful).

When tasking a computer with speaker recognition a complex set of technical processes must occur successfully to obtain a high degree of reliability. If we first consider a simplistic example of comparing two speakers, using an ASR system. Speech from both speakers will be converted from vibrations in the air into a digital recording. The speech from the ‘target’ voice in the recording is isolated, from any other speaker, through editing (diarisation). A feature extraction process is applied to the digital speech signal. The ASR system creates a statistical model of the speaker, considered to be reflective of the dimensions and geometry of his/her vocal tract, from the small section of speech supplied and in reference to samples of population data. Once the statistical models are created, complex pattern comparisons are undertaken between the model from the unknown speaker (often referred to as ‘test audio’ or questioned speaker) against the validated/enrolled speakers held in the ASR system (commonly referred to as ‘speaker models’ or known speaker samples’). This calculation pertains to similarity. Calculations are then also made in relation to the third set of normative data or population set to provide an estimation of typicality. Output is then provided as a probability, or likelihood ratio (LR) value, which may provide support for one of two hypotheses –

---

<sup>§</sup> Throughout this thesis the term automatic/automated speaker recognition (ASR) is used. Note that the acronym ASR is also frequently used to describe automatic/automated speech recognition.

(a) that the speech samples came from the same speaker (commonly referred to as  $H_0$ , or the null hypothesis), or (b) that they came from different speakers ( $H_1$ ). Both  $H_0$  and  $H_1$  must be tested.

Throughout the processes conducted, the operator of the ASR (or analyst/practitioner) often has a significant role to play such that it could be argued that they are effectively part of the system. For example, the operator will select which speech should be used for comparison and may determine through further editing what audio data should constitute the speaker model(s). This process will be applied to the questioned audio. In changing the settings an operator(s) may also influence the selection of the underlying normative data (the population). Furthermore, the operator may adapt the ASR based on their experience of optimising it. This may include changing how information is extracted from the speech (feature extraction), how the statistical modelling is completed or aspects such as calibration and threshold setting. Finally, the user is required to use their skills and experience to interpret the numerical LR (or log likelihood ratio) output and produce a report which can be understood by a non-specialist person.

A fundamental requirement is to separate the measurement of the speech signal from non-speech in the recording and therein lies an enduring problem. How to effectively separate the measurement of desirable variability (speech) whilst removing the undesirable variability of the recording and the end-to-end signal path, the ASR system and any additional variability or even bias that could be introduced by the operator. To fully understand the operating limits of the technology we need to measure audio/speech quality and understand what constitutes an acceptable amount of speech data at a high enough quality to produce an acceptable output. By doing so, we can decide when the influence of contaminants and inhibitors on ASR performance is significant enough to determine that ASRs cannot be meaningfully applied. Should we ever hope to transition ASRs to court use (forensic application) we need to better quantify the acoustic conditions in which ASRs work accurately and reliably and those in which they do not. These fundamental questions formed the motivations for the research conducted in this thesis.

## **1.2 Research Aims**

This thesis examines the influence of acoustic variability on ASR performance. Many variables affect the performance of an ASR system. These fall into two broad classes: inhibitors and contaminants. Five sets of research experiments are conducted examining the significance of speech quantity (inhibitors) and the technical quality of the audio recording (contaminants) on ASR performance. It will also be demonstrated that inhibitors and contaminators are linked variables, i.e. quality affecting quantity of net speech suitable for ASR comparison. The experiments in this thesis are therefore driven by three core objectives.

The first objective is to produce a comprehensive set of measurements pertaining to ASR performance under five commonly encountered acoustic conditions. The purpose of which is to assist with informing casework practitioners when applying ASR systems to case data. In addition, detailed metrics are provided to assist with determining the points at which acoustic degradation is likely to be too extensive to obtain meaningful ASR results (i.e. ASR system application would not be recommended). This objective acknowledges that ASR performance varies between systems which use different normative data and settings on variable case data.

The second objective is to examine two types of ASR systems and evaluate how performance differs with respect to acoustic variability. The purpose of assessing the difference in performance is to assist the casework practitioner in determining which types of ASR systems may demonstrate greater (or lesser) resilience to acoustic variability. It is also intended that the data from these experiments will be useful to casework analysts and those who design future ASR systems and/or integrate them across networks.

The third objective is to inform discussion and provide recommendations with regards to ASR suitability for forensic speaker comparison (FSC) within the context of acoustic variability. The purpose of this objective is to examine questions such as whether acoustic variability could prevent repeatable and reproducible ASR results and the extent to which generational improvements in ASR systems mitigate against acoustically degraded data.

### **1.3 Thesis outline**

Chapter 2 provides an overview of the literature that informed the thesis, inspired the research questions and guided the subsequent categories of experiments conducted. Additional literature reviews are provided at the beginning of each chapter pertinent to the specific subject areas.

An explanation of technical terms and concepts is provided in Chapter 3 'From Speaker Source to Analytical Destination' which follows the speech path from formation, through a typical audio path, culminating with the ASR system and practitioner.

Chapter 4 provides a summary of the research questions. Chapter 5 presents the methodology and materials common to the experiments conducted. Chapter 6 provides a summary overview of the preliminary tests and baseline experiments - these assisted with informing the methodology, defining the research direction and the scope of the experiments.

Chapters 7 to 11 provide detailed documentation of the 5 categories of experiments. These pertain to: speech quantity (net duration), SNR, reverberation, frequency bandwidth and transcoding or codec(s). Results and observations are then presented with a summary discussion.

Chapter 12 discusses all the experiments relative to the objectives of the thesis and examines the wider implications of acoustic variability on ASR usage. Recommendations are offered regarding the integration of ASR systems into speaker comparison work including the enduring issues relating to transitioning ASRs into the evidential process. Chapters 13 and 14 discuss opportunities for future research and conclude the thesis.



# Chapter 2 Literature Review

---

This chapter presents an overview of the literature that informed the research questions and thesis objectives. It also places the ASR experiments into the wider context of forensic and investigative application. Additional research reviews pertaining to each set of experiments conducted are presented within chapters 7 to 11.

## 2.1 Speaker Comparison Methodologies

In discussing either the investigative or forensic application of ASRs, definition is required as to the two main types of use cases. The European Network of Forensic Science Institutes (ENFSI) presents the following terms and these definitions are adopted for the purpose of this thesis.

**Forensic:** Seeking to establish facts of interest using science and technology in the context of the law or in a law court. ENFSI also refers to this as the ‘evaluative mode’ (2015: p.3).

**Investigation:** A systemic enquiry, examination, study and survey of facts, circumstances, situations, incidents and scenarios in order to render a conclusion. ENFSI refers to this as the ‘investigative mode’ (2015: p.3).

Gold and French (2011) surveyed 36 practitioners from 13 countries. In undertaking speaker comparison casework, five categories of methodologies were found to be used by experts. The common methods of analysis were described (2011: p.296) as:

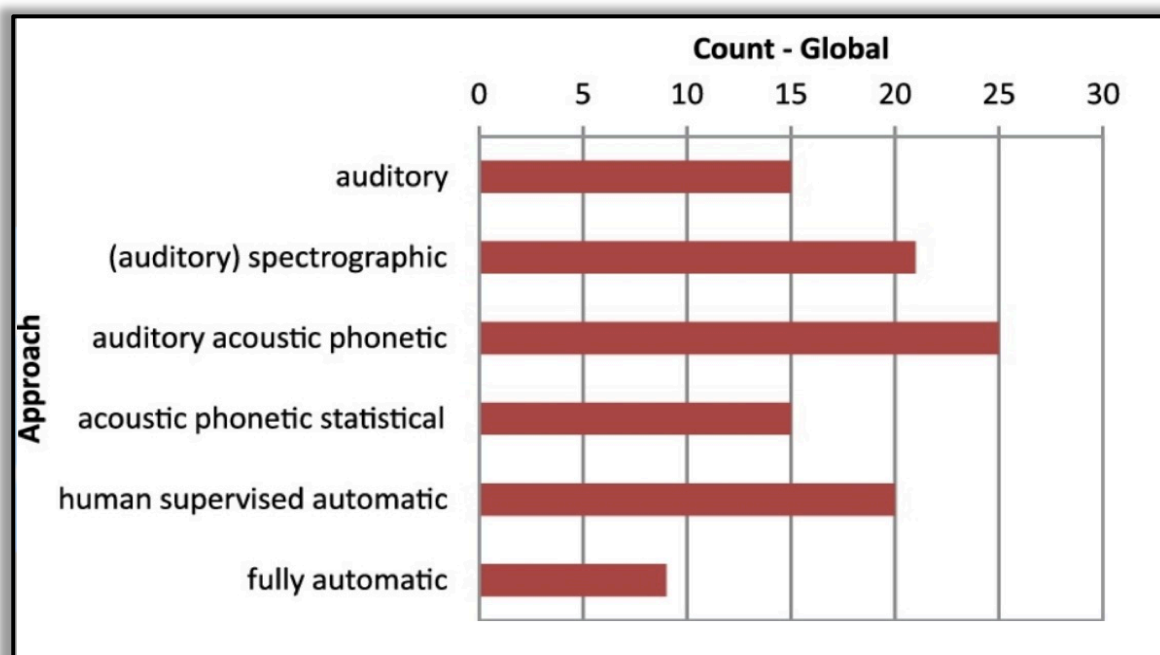
- i. **AuPA.** Auditory and Phonetic Analysis. Analysing speech through comparison of segmental and supra-segmental features
- ii. **AcPA.** Acoustic Phonetic Analysis. Quantifying physical parameters of the speech signal using analysis software.
- iii. **AuPA and AcPA.** Combination of Auditory Phonetic and Acoustic Phonetic Analysis
- iv. **ASRs.** Automatic Speaker Recognition system
- v. **HASR.** Automatic Speaker Recognition system (i.e. with human assistance).

Gold and French (2011) assessed that the methodology differed dependent on the organisation (research institute or university, government agency, private laboratory or individual) with 74.46% of organisations tending towards methodology 3. Only 17.02% used HASR methodology (5) and no organisations solely applied ASRs at that time. It was also noted that the majority of HASR users (33% in comparison to 16%) were found to predominantly reside in the government/law enforcement sector rather than universities and research institutes. Whilst explanation is not offered it is suggested

that the application of ASRs in investigative casework is likely due to the time-bound element and pressures of scale - i.e. potentially much larger quantities of speech data and very limited resources/time to complete detailed auditory assessments for multiple speakers.

A later international survey was undertaken by Morrison et al. (2016) on behalf of Interpol. This investigated the use of speaker identification, by international law enforcement agencies (LEAs) and received 91 responses from 69 countries (Figure 2.1). The group reported an upward trajectory in terms of ASR use (referred to as ‘human supervised automatic’ and ‘fully automatic’).

**Figure 2.1:** Morrison et al. (2016: p.94). Summary of LEA speaker comparison methodologies



Gold and French (2019) produced a second survey polling 39 forensic speech scientists. They too reported a rise in use of ASR systems with 41.2% of respondents using ASRs in comparison to 17.02%. Of those ASR users, 78.6% applied acceptance criteria in regards to technical quality. Whilst thresholds were not standardised across practitioners, users were broadly aware that audio quality variability can influence speaker comparison casework and completed technical assessments for acceptance. In relevance to this thesis the Gold and French (2019) survey noted acceptance criteria broadly defined as follows: minimum net duration for acceptability ranged from 3s to 20s, SNR from 10db to 25db, minimum high frequency values were 3.4kHz and 4kHz with minimum sample rate as 8kHz (2019: p.7). Their survey also noted that an average of 30% of submissions failed ASR acceptance criteria (2019: p.7).

French and Stevens (2013: p.5) pointed out that the advantages of using HASR/ASR systems include the reduction in subjectivity compared with other types of analysis (e.g. AuPA) in addition to the speed of ASR operation and the replicability of results. ASRs can clearly provide an opportunity to

produce a more empirical analytical method, offering the repeatable and reproducible output criteria as recommended by quality control standards such as the International Organisation for Standardisation (e.g. ISO17025). Furthermore, research by Campbell (2014) demonstrated that human assisted speaker recognition systems are starting to outperform analysts (although not expert practitioners) in NIST high confusability trials. Their research showed that 14 out of 15 HASR comparisons were found to be correct in comparison to 11 out of 15 (for naïve listeners). In light of all the research, however, it was important to note that all surveys reported that almost none of the practitioners relied solely on ASR output.

In summary, the practitioner surveys broadly show that ASR systems are becoming more widely used. This is likely due to factors such as the greater prevalence of underlying technology and the increase in communications methods such as voice over internet protocol (VoIP) and instant voice messaging (e.g. smart phone applications). The latter of which contributes to greater quantities of audio data events requiring analysis for which a human alone cannot process. In addition, ASR systems have been progressively improving in performance, which will be discussed in later chapters and this is likely building confidence and trust in output. Finally, ASR systems offer an opportunity to provide more objective measurements of analysis (repeatability and reproducibility).

## **2.2 Forensic Speaker Comparison**

In the U.K., forensic speaker comparison (FSC) refers to the process of conducting human/auditory analysis completed by an expert and then presented in court. Whilst objective acoustic measurements may also form part of the comparison process such as formant frequency, vowel and/or consonant measurements or voice onset time, ASR systems are not yet incorporated into the evidential chain in the UK. Broadly speaking, this is because ASR systems apply a different form of acoustic analysis, consisting of a feature extraction process and statistical modelling, in order to conduct a speaker comparison (in reference to a normative population). A detailed explanation of how ASR systems work is provided in Chapter 3.

Whilst certain countries do accept ASR systems into evidence, most ASR usage resides in the investigative domain as originally stated in Decker and Handler (1977). Although it is important to note that early pattern matching systems differ considerably from later GMM-UBM systems, with more complex feature extraction and the addition of normative data. In seeking to obtain explanation as to why ASRs are not admissible in U.K. courts, fulfilment of acceptance criteria often relates to the implementation of the scientific method, presentation of expected error rate and the capability and reliability of the expert witness. In the U.S. three cases refer to the admission of expert testimony referred to as the Daubert Standard (Daubert v. Merrell Dow Pharmaceuticals, 1993). The judge then determines whether those criteria are fulfilled and in U.K. courts these, similar standards, are clearly difficult to apply to ASR systems. In the U.K. a recent Court of Criminal Appeal (England

and Wales) case tested the admissibility of results from an ASR system into evidence in *R -v- Slade and Ors* [2015], EWCA, Crim 71. Professor Peter French and Dr. Phillip Harrison (JP French Associates), as U.K. expert witnesses, sought to apply ASR results in addition to auditory and acoustic phonetic evidence. The ASR analysis they conducted was completed using an Agnitio Batvox system (2009). The court would not accept the ASR results into evidence. To summarise, the court cited the unsuitability of population dataset, the potential lack of reproducibility of results across different ASR systems, likelihood output pertaining to small quantities of speech data (and a difficulty in interpreting statistical results) as significant factors (French, 2017: p.5) which all required addressing and prevented admissibility. Nonetheless, the test case was instrumental in questioning why ASR systems should not be used. The case also galvanised many in the U.K. forensic speech community to continue to progress ASR systems, processes and methodology to forensic application in the future and the case also influenced the research conducted in this thesis.

In a study completed by Morrison (2018b) pertaining to a 2017 case in New South Wales a similar conclusion was reached regarding population data and the application of a GMM-UBM ASR (open source). Morrison summarises with recommendations to use speech data which accurately reflects the conditions of the case (2018b: p.[e]6).

French and Stevens (2013: p.4) proposed that approximately 70% of the forensic casework that their company was tasked with was forensic speaker comparison (FSC), defined as:

‘...the comparison of a voice in a criminal recording with that of a known suspect, the purpose being to assist the courts with determining identity or non-identity of criminal and defendant’. (2013: p.4).

Whilst this process refers to auditory speaker comparison completed by an expert practitioner (rather than HASR or ASR approaches) an important aspect to note is that the process itself does not determine identity, which remains the domain of the courts. It is the expert who provides a view as to the strength of comparison for two competing hypothesis for same speaker (H0) and different speaker (H1). In addition, if intending to transition ASRs to forensic application there is also an associated risk with potentially transferring some of the responsibility for informing the courts with the strength of those hypotheses away from the expert(s) and towards ASR system(s).

The potential risks of transitioning ASR systems to forensic application when there are so many unknowns remain significant. Eriksson recommends that ‘a forensic speech expert knows the tools they use inside out’ (2012: p.48). However, this can be difficult to achieve on a complex system and given the high variability of speech and acoustic conditions. Expertise also requires diversification across many fields (e.g. linguistics, acoustics, signal processing, phonetics, mathematics, statistics and IT/engineering).

A position statement from the International Association of Forensic Phonetics and Acoustics (IAFPA) challenges the use of likelihood ratio approaches where population data is not available for reference when conducting FSC casework (2007: p.5) and this debate extends to the significance of normative data when applied to ASR systems. In the context of human conducted auditory speaker comparison the ‘UK position statement’ French and Harrison (2007) recommended that the term ‘forensic speaker comparison’ should replace ‘forensic speaker identification’ (2007: p.8) as a likelihood ratio is preferable to a binary decision. Prior to the position statement and in the context of auditory phonetic analysis Nolan (1983) had asked how reliably individuals can be recognised by voice at all, arguing that the plasticity of the vocal tract and variability between speakers is not fully known and measured. In summary, speaker comparison clearly cannot produce a definitive match/non-match output yet, arguably, humans tend to expect computers and by extension ASR systems to produce binary decision outputs.

Rose and Morrison (2009) also agreed in their response to the position statement stating that ‘identification’ and ‘recognition’ could carry the connotation of an absolute conclusion or a posterior decision (2009: p.146). On a related point, the statement and response also agree in the importance of distinction between the likelihood of the evidence given the hypothesis (province of the forensic scientist) and the likelihood of the hypothesis given the evidence (province of the court). Further explanation of this is presented in chapter 3 (Bayes theorem and likelihood ratio calculations).

## **2.3 Automatic Speaker Recognition Systems**

In an early review of automatic speaker recognition systems conducted for the institute of electrical and electronics engineers (IEEE), Rosenberg (1976) summarised the research and development work conducted to date. Pattern matching systems were then the prevailing technology but were prohibitively expensive due to computing costs. They were also predominately used by telecommunications companies and research universities who could afford the IT.

In the U.S. the Bell telecommunication laboratories pioneered early ASR development work building on the enabling technical advances made earlier in the 20<sup>th</sup> century (telecommunications, radio, recording and analogue to digital conversion). Early engineers who progressed speaker comparison from traditional auditory methodologies to pattern matching systems were Doddington (1970), Bricker and Pruzansky (1971) and Atal and Hanauer (1971). The early research systems developed for modern telephone speaker recognition and entry control systems were further progressed by teams at Texas Instruments. Rosenberg’s study (1976) made three important points which influenced the research questions and subsequent experiments conducted in this thesis.

“Factors which can loom large over the implementation of a speaker-recognition system are the recording environment and the conditions governing the transmission of the speech signal to the processor.” Rosenberg (1976: p.479).

“In the design of such systems, careful allowance should be given to the effects of background noise and room reverberation at the source and the reduced bandwidth, distortion, and line disturbances.” Rosenberg (1976: p.480).

“Most evaluations have been carried out in the hothouse atmosphere of the sound booth and high-quality recordings. Eventually, however, one must consider whether these conditions represent a fair approximation to conditions that are expected in a practical application.” Rosenberg (1976: p.479).

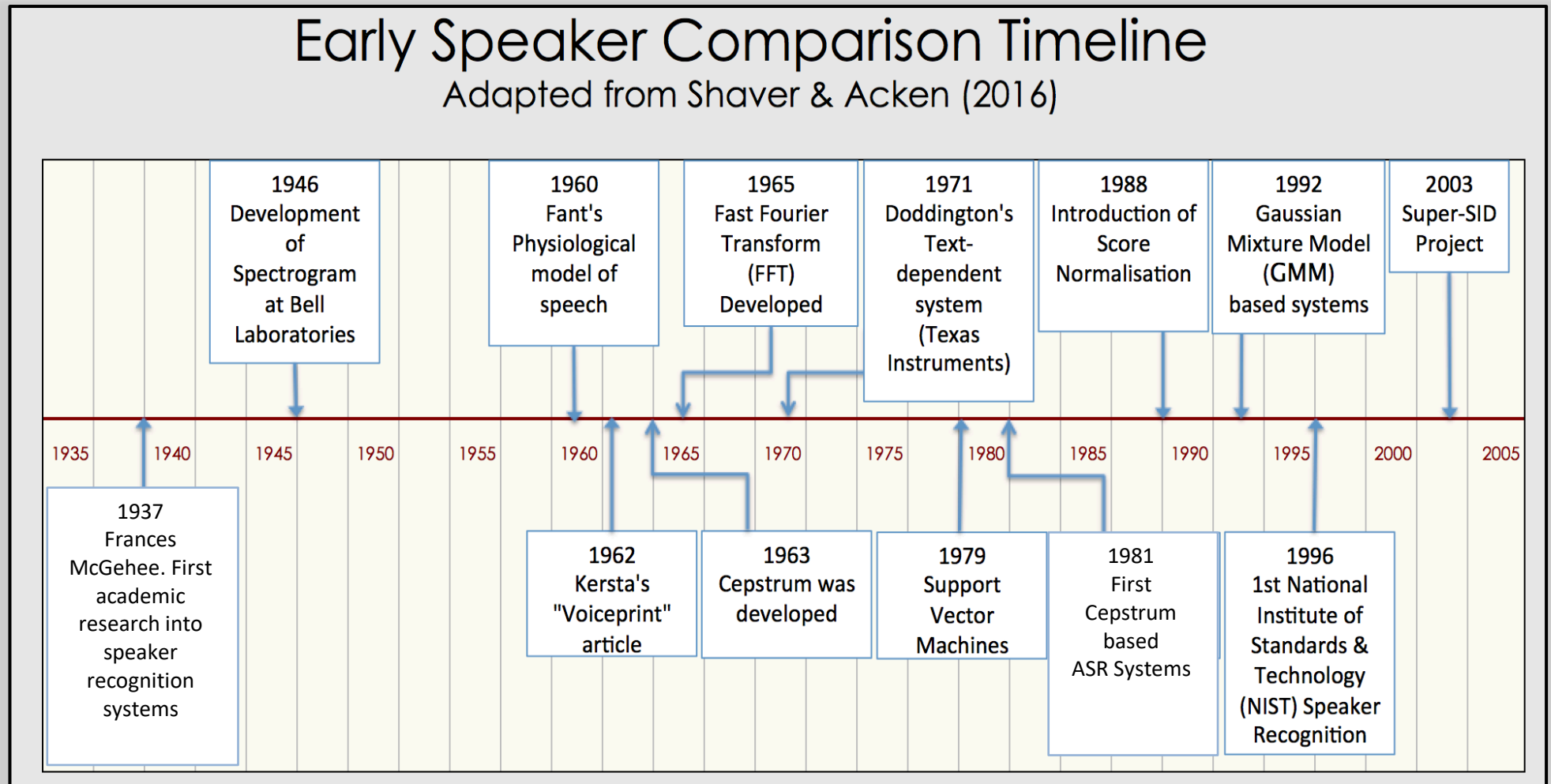
In the late 1970s and early 1980's the application of speaker comparison systems for law enforcement purposes was researched by the Phillips laboratory in Germany led by Bunge (1976). This research, part government sponsored, used the AUROS (AUtomatic Recognition Of Speakers) corpora to test an early acoustic pattern matching system. The AUROS database was documented as containing 5,000 utterances (which) were the same apart from the name of the speaker given in each utterance. Bunge's pattern system used a 43 channel filter bank to capture information into a classifier from the long-term-averaged spectra (100Hz to 6kHz) at 50 times per second. Whilst the Mahalanobis classifier, developed in 1936, was relatively rudimentary in comparison to later classifiers (using standard deviation calculations from extracted mean values) the results from early experiments were extremely encouraging. Bunge demonstrated that 2,500 utterances could be correctly verified for 100 speakers (82 Males and 18 Females) to an accuracy of approximately 99.5% correct verification (using the other 2,500 for enrolment). This impressive level of high accuracy for the time was likely assisted by: the use of identical utterances (in terms of content) that existed in both the input and database i.e. semi-text dependent application, the limited quantity of utterances per speaker (between 10-20) and a high degree of channel matching between questioned and enrolled utterances. Bunge's early research suggested that the frequency bandwidth of the telephone channel was a factor hampering recognition performance (1976: p.206). Nevertheless, the potential of ASRs to assist with speaker comparison in the future was noted and Bunge proposed potential use cases to assist with verifying the speaker identity of criminals such as 'blackmailers and kidnappers' (1976: p.207).

In other research, Wolf (1972) demonstrated pattern classifying experiments using 6 read sentences from 21 male speakers aged from 22 to 42. This also provided a rudimentary but effective speaker verification system with an error rate of 2% and was seen as a substantial step forward. The early pattern matching programs and variations such as Doddington's Texas Instruments system (from the 1970s and 1980s) provided influence to Rosenberg, Lee and Soong (1990). They produced a research system for AT&T/Bell speech research laboratories able to identify 100 speakers (50 male and 50

female) on a corpus containing 20,000 digit utterances. The utterances were band filtered (200Hz to 3.2kHz) and accuracy was assessed to be 7% to 8% equal error rate (EER) on a single digit test utterance (0.5s per digit) and less than >1% on 7 digits. This was an impressively low error rate albeit on text dependent (identical digit utterance) verification. The group cautioned that a higher EER% for text independent and larger vocabulary use would be expected (1972: p.269).

In the 1980s and throughout the 1990s computers progressed to become faster, cheaper and therefore more ubiquitous. In the early 1990's Rose and Reynolds from MIT/Lincoln Labs used the improvements in statistical modelling (gaussian mixture models or GMMs) to generate speaker models from speech files. Reynolds (1994) further improved on GMM speaker verification methodology and progressed the accuracy of systems with the inclusion of normative data, citing Higgins, Bahler and Porter (1991). This was significant because systems evolved from pattern matching systems, determining similarity between files, to considering typicality against population (or normative) data. Reynolds also incorporated likelihood ratios for presenting output and this effectively gave rise to the modern ASR comparison system. Saquib et al. (2010) supported Bill Gates' view from the late 1990's that voice biometrics was becoming one of the most important IT innovations of the time. Shaver and Acken (2016) produced a summary of the early advances that contributed to modern speaker verification systems. This is adapted and reproduced in Figure 2.2 (next page).

Figure 2.2: Shaver and Acken (2016) early speaker comparison timeline (reproduced)





Throughout system evolution, the ideal conditions for a comparison system to operate successfully were also scrutinised and documented by Wolf (1972: pp.2044-2045), Nolan (1983) and Eriksson (2012: p.58). These are summarised in a consolidated list (Figure 2.3) and are widely accepted as the fundamental requirements which underpins ASR methodology - in addition to other auditory and acoustic speaker comparison methods.

**Figure 2.3:** Summary of ideal speaker comparison conditions

<u><b>Ideal conditions for successful speaker comparison</b></u>	
1)	Large between-speaker and small within-speaker variability. <b>W, N</b>
2)	Be difficult to impersonate/mimic. <b>W, N</b>
3)	Not be affected by the speaker's health or long-term variations in voice. <b>W</b>
4)	Occur frequently and naturally in speech. <b>W</b>
5)	Be easily measurable. <b>W</b> ' <i>Measurability and Availability.</i> ' <b>N</b>
6)	Not be affected by background noise nor depend on the specific transmission medium. ' <i>Robustness in transmission</i> '. <b>N</b>
7)	Occur naturally and frequently in speech. <b>W</b>

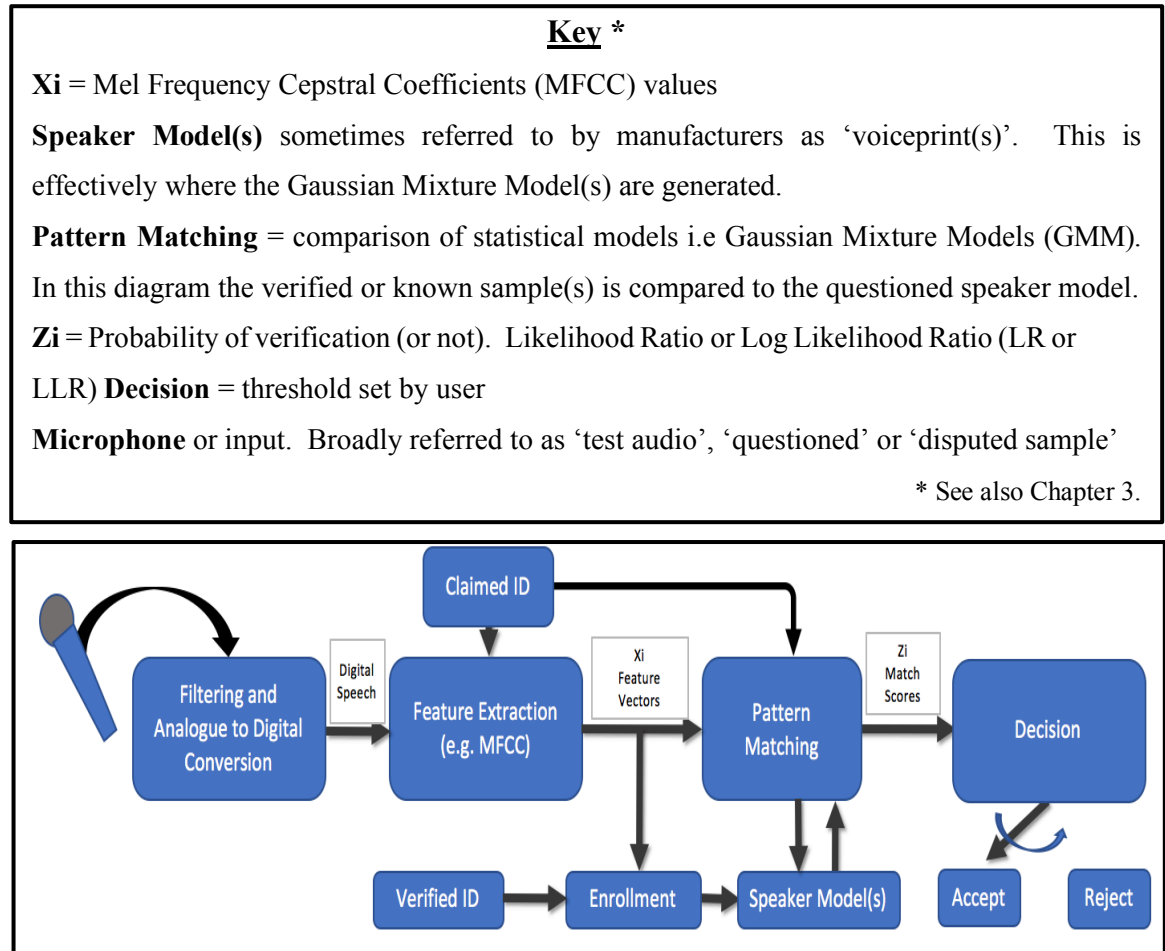
**W:** Wolf   **N:** Nolan

These are widely considered as ideal conditions, however even the pioneers of early systems quickly realised that to fulfil all those criteria and obtain the perfect conditions was not possible at the time (Wolf, 1972). The points from Figure 2.3 were applied to the experiments completed in this thesis. It was noted that, whilst conditions were broadly satisfied, they could not be considered completely 'ideal'.

For example, the control corpus used (DyViS) consists of males between a small age range (18-25) and within speaker variability was constrained by session data variety (2 types of conversational speech). Neither points 2 nor 3 applied to the experiments since the control corpus did not contain impersonations and the recordings were completed over a short time frame for which neither long-term variations (nor health) applied. The recordings were recorded under controlled conditions and of sufficient (high) quality, such that measurements could be extracted (point 5). Degradation was applied under controlled conditions (point 6), however it was noted that some recordings were affected by the specific transmission medium (e.g. telephony codec) and noise, although negligible, was not completely absent (e.g. speaker 012). Nevertheless, since these factors were consistent in all baseline measurements it was determined as acceptable data for experimentation. Finally, re point 7, the conversations were somewhat staged in terms of content but flowed freely and were not read. The fake place names, used by all participants, were artificially constructed to be phonetically rich and varied in nature. It is suggested that this may have assisted with balancing (phonetic) content across speakers which, whilst not necessarily realistic, was consistent across all speakers at baseline.

ASR systems verify speaker identity through software comparison of a questioned sample and a known sample of speech (Campbell, 1997). The simplified flowchart below – Figure 2.4 adapted from Campbell (1997) - provides explanation as to how the early ASR systems operated.

**Figure 2.4:** ASR process, reproduced and adapted from Campbell (1997: p.1438)



Whilst the underlying process has not changed significantly, reference is not made to normative data (Figure 2.4). The later addition of normative data provided performance improvement by incorporating additional statistical distance measurements between question/test audio and normative data and also speaker model and normative data. Differences between question/test audio and speaker model reflect similarity but the addition of normative data provides additional measurement as to typicality. This methodology informs the calculations required for likelihood calculation (see 3.5.1).

Research undertaken by Becker, Solewicz, Jardine and Gfroerer (2012) in applying ASRs to actual case data was influential to this thesis. The group completed multiple ASR experiments using a new GFS1.0 corpus containing recordings of male German speakers recorded in case/forensic conditions (39 offenders and 21 suspects). These were taken from the German Federal Criminal Police Office Bundeskriminalamt (BKA) case files. Their study was innovative in utilising real case data where the correct outcome was effectively known, as far as was practicable. The team examined the

performance of 7 ASR systems deployed in Israel, France and Germany. The systems varied in architecture. The two Israeli and one unnamed commercial systems utilised a very early i-vector statistical modelling architecture. Results showed EER ranges from 9% to 12% with the i-vector systems marginally outperforming the GMM-UBM systems. Whilst these scores demonstrated relatively good performance they were notably poorer than comparable ASR systems tested on high quality/test corpora. The group also noted that ASR performance was particularly degraded for those recordings that used a handheld pocket audio recorder (Dictaphone) in the signal path. Their research conclusions are summarised below (Figure 2.5).

**Figure 2.5:** Key conclusion points reproduced from Becker et al. (2012: pp.5-6).

- i. Using automatic forensic voice comparison systems without any further investigation of the recording material results in a considerable proportion of errors.
- ii. The recording device properties with or without the transmission channel influence seem to affect automatic systems severely.
- iii. Because of system sensitivity to recording and transmission channels, auditory and acoustic evaluations of the channel properties are mandatory.
- iv. Automatic voice comparison systems do not account for linguistic features such as dialect, accent, sociolect etc. When there exists strong contrary evidence from forensic phonetics and automatic systems, the expert has to decide which evidence is more reliable.
- v. Automatic voice comparison systems are based on acoustic properties of speakers and are generally assumed to be language independent. However, in cases where the user does not have a thorough knowledge of the language in question, an exclusion of errors based on linguistic analysis is impossible.

The team also compared the Dictaphone material results against the non-Dictaphone results and suggested that the influence of frequency bandwidth and/or data compression (codec) could also influence ASR performance. Their paper concluded by recommending further research to collect more data and the development of better guidelines as to when ASR systems should be used (or not). The output of their research paper therefore assisted with informing the experiments conducted in this thesis.

# Chapter 3 From Speaker Source to Analytical Destination

---

This chapter explains the technical concepts discussed throughout the thesis and is structured to follow a typical end-to-end process from speech production, through a typical signal path, to the ASR system and analyst practitioner. It begins with the formation of speech, progressing through the digital recording process/signal chain and culminating with the ASR system and interpretation of results.

- 3.1 Speech Production
- 3.2 Intrinsic and extrinsic variability
- 3.3 Audio recording and the signal path
- 3.4 ASR systems
- 3.5 ASR output (LR, LLR) and performance measurement (FAR, FRR, EER, Cllr)
- 3.6 ASR use cases
- 3.7 Summary

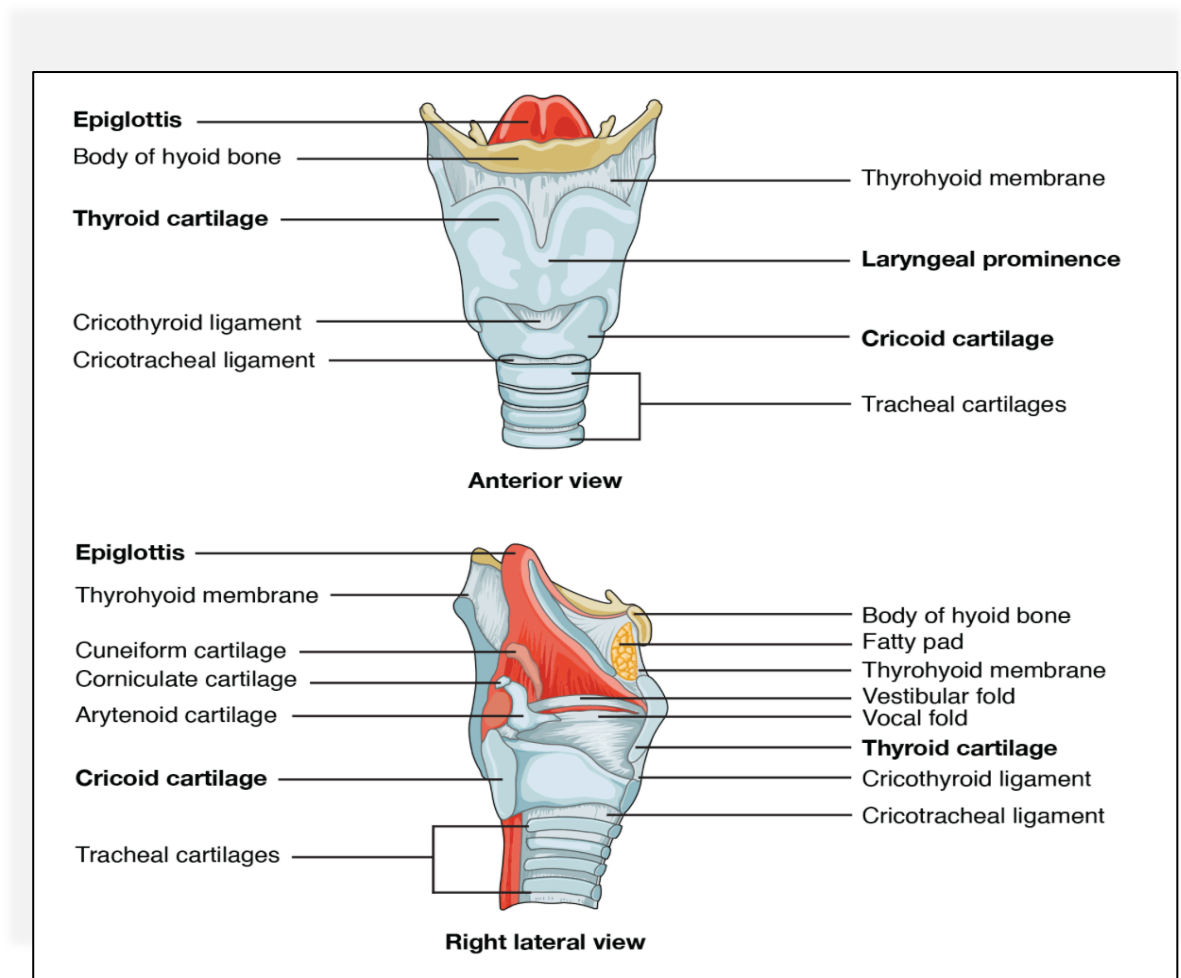
The above topics are highly complex and a degree of simplification is therefore unavoidable. This chapter is limited in scope to provide a foundation explanation of relevant terms only. Complex or unusual audio capture methods are not referenced and the recording/signal path described is intentionally pertinent both to the corpora used in this thesis and typical of audio files generally presented for casework analysis (i.e. telephone and interview/room recordings). References are also provided throughout for further reading. The chapter concludes with a summary discussing the complexity of the end-to-end process and inherent variability therein.

## 3.1 Speech Production

Fry (1979) describes the sound waves of speech as amongst the most complex in nature. When analysing, or measuring, speech sounds, it should be noted that not all are voiced (with vocal fold vibration in the larynx). For example, the sounds in English /ch, f, h, k, p, s, sh, t and θ/ are formed unvoiced (with no such vocal fold vibration and wide opening of the glottis, i.e. the space between the vocal folds). Most sounds are driven by a pulmonic egressive airstream mechanism, i.e. using airstream modulation (i.e. tongue, lips, teeth and jaw) drawn from the lungs (pulmonic). Some languages, e.g. Damin or Bantu, use other types of sound generation such as tongue-based or bilabial clicks and, in some Scandinavian languages, some speech sounds are made through pulmonic ingressive breathing (inhaling) - see also Ekland, (2007; 2015).

This is important because, with variability in speech sound creation, it should therefore not be assumed that an ASR developed and tested on one language will necessarily work to exactly the same efficacy on all languages – or mix of languages. For the purposes of the experiments completed in this thesis all speech data, including the normative sets, are English language. Voiced speech has its source within the larynx, a structure formed of cartilage that encloses two strips of muscle known as the ‘vocal folds’. The larynx (Figure 3.1) provides a ‘vocal note’ (Greene and Mathieson, 2001: p.5) and requires modification to form speech.

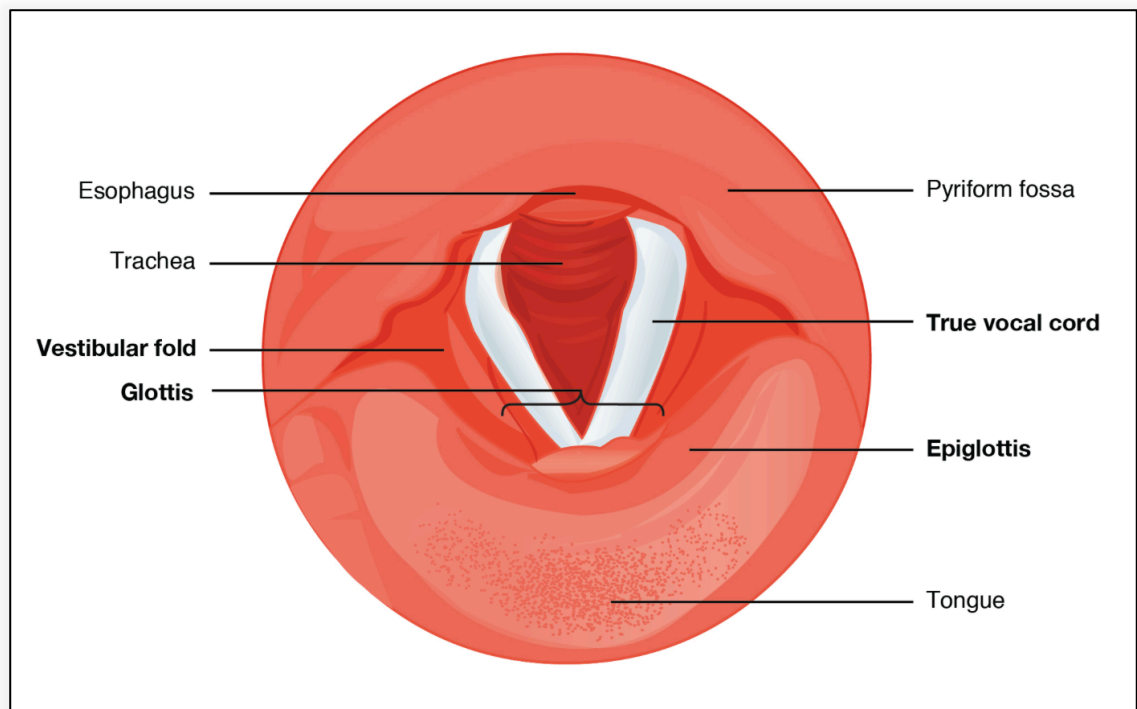
**Figure 3.1:** Construction of the larynx. Anterior and lateral views  
From: Opentextbc.ca



During the act of speaking, the vocal folds (Figure 3.2) are generally in one of two main positions. First, they may be held apart (i.e. open position), allowing the unimpeded passage of egressive air from the lungs, as for the consonant sounds that one terms as ‘voiceless’. Second, they may be held in loose contact, such that when the exiting air passes between them they vibrate, producing the effect known as ‘voicing’, which characterises a further set of consonants and all vowels. Voicing, or phonation, at the laryngeal source consists of a relatively ‘pure’ note. Perceptually, this is referred to as ‘voice pitch’, and may be experienced as high or low or anywhere in between. The opening and closing of the larynx is referred to as glottal pulses. Glottal pulses are visible using external analysis software such as spectrograms, discussed later in this chapter.

**Figure 3.2:** Cross section of the vocal folds (and glottis)

From: Opentextbc.ca



Acoustically, voice pitch can be measured – in fact, estimated – and computed in terms of Hertz (Hz)\*\*. These are vibratory cycles of the vocal folds (per second), referred to as ‘fundamental frequency’, F0 or the ‘vocal note’ (Fry 1979: p.65). While the rate of vocal fold vibration is constantly varying throughout the act of speaking, it can be averaged over any stretch of speech, providing an average F0 value. For women and young children this tends to be higher than for men owing to the fact that men generally have more flaccid vocal folds that are longer, heavier and of greater mass.

Research by Hahn et al. (2006: p.1104) stated that the average vocal folds, for adults, are approximately 10-15mm in length and 3-5mm in thickness. Alternating between flaccid and tense vocal folds alters F0 and provides intonation. With respect to auditory perception, pitch does not necessarily equate directly to acoustic measurements. Nevertheless, Fry (1979) estimated that for male speakers the average F0 is approximately 120Hz and for women approximately 225Hz. Children generally have a higher average mean F0 at approximately 265Hz (Fry 1979: p.68).

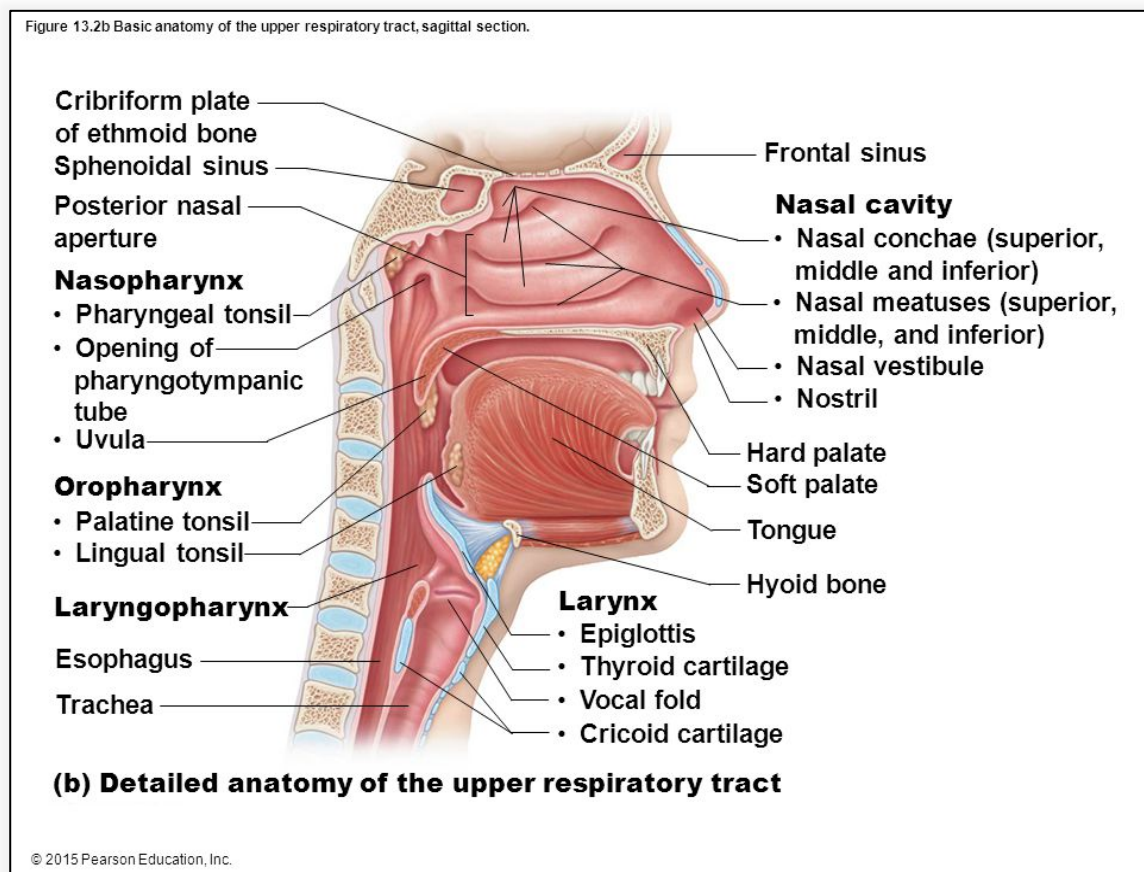
From the larynx, the vibrating air passes into the vocal tract comprising the pharynx, the oral cavity, and, for some sounds, the nasal cavity. As it passes through these supralaryngeal resonating chambers, they act as filters, shaping its energy-frequency content. Specifically, harmonics of the

---

\*\* Frequency is described in Hertz (Hz) named after the physicist Heinrich Hertz 1857-1894 and refers to the numbers of wave cycles per second.

F0 and the areas surrounding them are amplified or dampened according to the shape and size of the resonating chambers and the disposition of the articulatory organs, tongue, velum and lips (Figure 3.3).

**Figure 3.3:** Detailed sagittal section of the respiratory tract  
From Pearson education: [slideplayer.com/slide/4876905/slide6](http://slideplayer.com/slide/4876905/slide6)



With vowel sounds, harmonic areas that are amplified are known as ‘formants’. Fry (1979) considered formants to be ‘the true resonances of the vocal tract’ (1979: p.78) and effectively the basic building blocks of speech. Formants are referred to as F1, F2, F3 and F4 as they extend upwards in frequency from low to high and can be viewed as energy bands on spectrograms (in the horizontal plane).

It is the formant structure of vowels that provides them with their distinctive quality and differentiates one vowel class from another. However, since individual speakers have somewhat different vocal tract dimensions and configurations, the formant structures serve not only to distinguish, say, /i:/ from /u:/ from /e/ etc., but also to differentiate different speakers’ productions of those vowels from those of other speakers (LaRiviere, Winitz and Herriman, 1975). Also, at the longer term, suprasegmental level, because speakers have different vocal tract settings, they are therefore also distinguishable from one another in terms of voice quality (see below).



### 3.1.1 Voice Quality

Voice quality (VQ) is described as those characteristics that are present more or less all the time that a person is speaking (Abercrombie, 1967). Voice quality is relevant to ASR analysis because the analytical units that ASR systems operate with, Mel Frequency Cepstral Coefficients (explained in 3.4.3) reflect a range of the major components of voice quality, namely those arising from vocal tract settings. This was recently examined in Hughes et al. (2017b) where their study confirmed that speakers with common supralaryngeal VQ profiles (in the context of 100 male speakers<sup>††</sup>) produced weaker ASR output i.e. lower true positive and higher true negative likelihood ratio scores. Kreiman, Vanlancker-Sidtis and Gerratt (2003) and Pisoni and Remez (2004) referred to one of the first descriptions of voice quality occurring in history as early as the 2nd century AD by Julius Pollux. Pisoni and Remez (2004: p.347) provided a Table of those original descriptors and those later added by Moore and Gelfer. This is reproduced in Table 3.4 below and the evolution of the descriptors is clear when compared with later profile analysis tables, which are more objective and relate to settings (Table 3.5).

**Table 3.4:** Pisoni and Remez (2004: p.347) early evolution of VQ description

<i>After Julius Pollux, second century AD<sup>a</sup></i>	<i>Moore, 1964<sup>b</sup></i>	<i>Gelfer, 1988</i>
High ( <i>altam</i> )	–	High
Powerful ( <i>excelsam</i> )	Ringing	Strong, intense, loud
Clear ( <i>claram</i> )	Clear, light, white	Clear
Extensive ( <i>latam</i> )	Rich	Full
Deep ( <i>gravam</i> )	Deep	Resonant, low
Brilliant ( <i>splendidam</i> )	Bright, brilliant	Bright, vibrant
Pure ( <i>mundatam</i> )	–	–
Smooth ( <i>suavam</i> )	Cool, smooth, velvety	Smooth
Sweet ( <i>dulcem</i> )	–	–
Attractive ( <i>illecebrosam</i> )	Pleasing	Pleasant
Melodious, cultivated ( <i>exquisitam</i> )	Mellow	Mellow, musical
Persuasive ( <i>persuasibilem</i> )	–	–
Engaging, tractable ( <i>pellacem, tractabilem</i> )	Open, warm	Easy, relaxed
Flexible ( <i>flexilem</i> )	–	Well-modulated
Executive ( <i>volubilem</i> )	–	Efficient
Sonorous, harmonious ( <i>stridulam</i> )	Chesty, golden, harmonious, orotund, round, pectoral	Balanced, open
Distinct ( <i>manifestam</i> )	–	–
Perspicuous, articulate ( <i>perspicuam</i> )	–	–
Obscure ( <i>nigram</i> )	Dark, guttural, throaty	Husky, guttural, throaty
Dull ( <i>fuscam</i> )	Dead, dull, heavy	Dull, heavy, thick
Unpleasing ( <i>injucundam</i> )	–	Unpleasant
Small, feeble ( <i>exilem, pusillam</i> )	Breathy	Breathy, soft, babyish
Thin ( <i>angustam</i> )	Constricted, heady, pinched, reedy, shallow, thin	Thin
Faint ( <i>difficilem auditu, molestam</i> )	Whispery	Weak
Hollow, indistinct ( <i>suburdam, obscuram</i> )	Covered, hollow	Muffled
Confused ( <i>confusam</i> )	–	–
Discordant ( <i>absonam</i> )	Blatany, whiney	Strident, whining
Unharmonious, uncultivated ( <i>inconcinnam, neglectam</i> )	Coarse, crude	Coarse, gruff
Unattractive, unmanageable ( <i>intractabilem</i> )	–	Shaky

<sup>††</sup> Dynamic Variability in Speech (DyViS) – see 5.3



Schemata were evolved in the 1960s by Voiers (1961; 1964) and Isshiki (Kreiman, Vanlancker-Sidtis and Gerratt, 2003). Honikman (1964) completed research into articulatory settings across several languages and incorporated specific references to the settings of the jaws, lips, and tongue affecting voice quality. Laver (1980) provided a more defined set of criteria and effectively established modern vocal profile analysis (VPA) based on the work of Abercrombie (1967). Laver evolved VPA with Mackenzie, Wirz and Hiller (1981) which progressed to form the modern VPA schemata used today by auditory experts. An example of a modern, full VPA Table from San Segundo and Mompean (2017) is below (Table 3.5) as based on Beck (2007).

**Table 3.5:** San Segundo and Mompean (2017: p.644) VPA Template

	First Pass		Second Pass						
	Neutral	Non-Neutral	Setting	Moderate			Extreme		
				1	2	3	4	5	6
<b>A. Vocal tract features</b>									
1. Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimized range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimized range						
3. Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
5. Pharyngeal			Minimized range						
			Pharyngeal constriction						
6. Velopharyngeal			Pharyngeal expansion						
			Audible nasal escape						
			Nasal						
7. Larynx height			Denasal						
			Raised larynx						
Lowered larynx									
<b>B. Overall muscular tension</b>									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						
<b>C. Phonation features</b>									
	Setting	Present		Scalar Degree					
		Neutral	Non-Neutral	Moderate			Extreme		
				1	2	3	4	5	6
10. Voicing type	Voice								
	Falsetto								
	Creak								
	Creaky								
11. Laryngeal frication	Whisper								
	Whispery								
12. Laryngeal irregularity	Harsh								
	Tremor								

VPA requires a trained practitioner to listen to speech and make judgments as to the vocal settings required to generate it. Kreiman and Gerratt (2000) stated that voice quality analysis should be

considered perceptual since it is somewhat reliant on the subjective opinion of a listener rather than objective measurements. Trials completed by Watt and Burns (2012) using lay listeners demonstrated that descriptions were often inaccurate or inconsistent, highlighting the requirement for trained practitioners. Nevertheless, there is a risk of inconsistency between experienced practitioners who disagree (Kreiman and Gerratt, 2000). Kreiman, Gerratt, and Ito (2007) further stated that accurate, replicable and valid assessments were difficult because all listeners have varying definitions of modal voice quality. In addition, the methods used to assess voice quality vary i.e. listeners differ in their own methodologies, their personal/mental representation of the population and interpretation of the task in hand. However, a significant empirical exercise and proposal for overcoming inter-rater variation through calibration is presented in San Segundo et al. (2018).

As discussed, most speech production involves a source (the larynx) and a filter (the supralaryngeal vocal tract). Assessment of voice quality is therefore broadly split in to two categories of features or settings, the phonatory and the supralaryngeal. Supralaryngeal features refer to those shaped by the tongue, teeth, lips, nasality, jaw settings and the raising or lowering of the larynx. Phonatory features reference those voice quality settings related to the creation of speech sound in the larynx ‘the production of voice at the glottal opening’ (Esling 2013: p.110). It is important to note that factors aside from physiology can influence voice quality. Esling states that certain languages, for example Swedish, have preferred long-term voice quality settings which can influence phonatory settings such as creaky voice (Esling, 2013: p.124). In addition, voice quality can also be influenced by social factors. Scherer and Giles (1980) found a correlation between social background and voice quality with a study which found separation between higher status ‘creak’ and lower status ‘whisper/harshness’. Voice quality assessment is of interest to both forensic practitioners as well as other sectors (e.g. medical) and, as Laver stated (1980: p.2), many voice quality characteristics are founded in the physiology of the speaker. A pilot study is currently underway by Gully et al. (2019) examining articulatory settings using MRI in closer detail.

Whilst it could be argued that vocal profile analysis makes a subjective auditory assessment there have been links noted between vocal profile analysis completed by trained experts and acoustic measurements, as noted by French and Stevens (2013: p.192) and Cardoso et al. (2019). For example, phonatory judgements such as creaky voice and breathy voice are related to the glottal pulse measurement. Supralaryngeal settings and the raising and lowering of the larynx are likely to influence formant measurement as found in Laver (1979). Stevens and French (2013) examined the voice quality of the 100 DyViS speakers used throughout the research experiments conducted in this thesis. Their analysis adapted previous methods to assess voice quality as completed by Abercrombie (1967) and Laver (1979). In summary, each speaker was scored using a subjective six-point scale (0 to 5) for 34 vocal settings. In the preliminary tests in this thesis, voice quality data was then examined in relation to the ASR output and (see chapter 6).

VQ features (VPA) can be extremely useful in comparing and discriminating speakers. Indeed, in an international survey of forensic speech scientists Gold and French (2012) found that most practitioners rated it the most important feature in forensic speaker comparison testing. However, completing VPA is time consuming, relies on a subject matter expert and is therefore not particularly practicable for applying to large volumes of speech files, for example. Acoustic measurements provide the best opportunity to transition from subjective assessment to objective analysis based on mean values from a population. If used by a trained operator, sympathetic to the complexity of vocal profile analysis, it is therefore suggested that there should be a place for semi-automated systems to potentially assist in this process and provide speed and scale.

### **3.1.2 Sound Pressure Levels**

To provide context to the signal to noise ratio (SNR) experiments conducted, a brief explanation of amplitude and sound pressure levels follows.

The volume of sound is measured in decibels (dB) and, effectively, the larger the number in decibels the louder the sound. Decibels are measured on a relative and logarithmic scale and a sound that is more than 10 times louder is referred to as 10dB whilst a sound which is 100 times louder is referred to as 20dB and so on. The sound pressure or volume of speech is important in the application of ASRs or auditory comparison work because speech sounds continuously vary with regard to the formant frequencies produced and the volume when they are formed.

In research pertaining to hearing loss Fant (1959: p.4) produced a summarised graph showing male speech (Swedish) captured at a distance of 1m. In plotting the frequency of formants F1, F4 and the fundamental frequency (F0) against sound pressure for vowel and consonant sounds (Figure 3.6) the variation of volume dependent on the speech sounds produced becomes apparent. Softer and unvoiced sounds at a lower sound pressure level are evident and to fully capture these sounds and measure them effectively a high-quality audio recording at close proximity with very low environmental noise is required or the SNR (See 3.3.6) will likely be poor.



**Table 3.7:** Examples of intrinsic speaker variability

Intrinsic	Examples	Comments
Gender	Fundamental frequency/pitch.	Previously considered binary (M or F), gender is increasingly viewed as a spectrum.
Age	Baby, child, teen, adult, senior. Laryngeal maturity/degradation.	Pitch, shimmer/jitter. Other factors influencing articulators – e.g. loss of teeth, surgery.
Language	Dialect, accent, code switching, slang terms and colloquialisms	Highly complex. Influence on ASR not fully understood.
Articulation	Precision of pronunciation (e.g. mumbling) – can influence vocal effort	Can therefore influence SNR (of recording).
Physiological	Chronic or temporal. Sickness, speech impediment, damage to larynx and/or articulators (teeth, tongue, mouth, nasal cavity).	Congenital, temporal (cold), chronic, changing (damage).
Style	Formal, familiarity, mirroring, declamatory, conversational, deceptive, conspiratorial, read speech. The interlocutor (conversational partner) can also influence speaking style.	
Emotion/behavioural	Anxiety, anger, depression, boredom, crying, excitable.	Currently in research space. Very difficult to effectively measure without significant baseline data.
Vocal effort	Raised/lowered volume, conspiratorial speech/whispering.	Can influence SNR of recording unless mitigated by closer proximity.
Temporal influence (drink/drugs)	Slurred speech, word choice, prosodic rate, dynamic range.	
Intrinsic/Extrinsic Physical obstruction	Balaclava, crash helmet, hand in front of mouth.	Can restrict movement of articulators in addition to dynamic/frequency range.
Duration of speech	Uncooperative, monosyllabic.	Multiple instances of monosyllabic responses into a speaker model can inhibit intra-variability.
Physical movement	Running, walking quickly, climbing stairs.	
Voice Quality	See also physiological, social factors.	See Chapter 3.
Disfluency/Filled pauses	Conversational hesitancy e.g. /erm, um, er/.	Considered useful for vowel data/measurement.
Repetitive filler words & phrases	E.g. ‘do you know what I mean?’	Could repetitive phrases influence ASR output? Combine ASR speech to text?
Sociolinguistics	Use of colloquial language, regional terms.	
Stress	Relaxed, uptight/formal, argumentative.	Behavioural/Temporal.

**Table 3.8:** Examples of extrinsic speaker variability

Extrinsic	Examples	Comments
Environmental noise	Wind buffeting, rain, thunder, hail, even temperature can affect.	Additive noise. Often outside of any form of control. Mic capture/position can mitigate, to some extent.
Net speech duration (Chapter 7)	Total length of speech extracted. (Also intrinsic if monosyllabic/non responsive).	Truncated recording. Proximity changes. Net speech is a significant factor, though with diminishing returns. See also intrinsic category (monosyllabic, lack of engagement).
Reverberation (Chapter 9)	Room reflections.	Can smear the speech at a sub second level. Mic capture/position and proximity can partially mitigate.
Machine noise/SNR (Chapter 8)	Vehicles, air conditioners, machinery/SNR.	Additive noise. Audio be enhanced, to some extent if fixed frequency/predictive.
Media noise	Television, radio, internet.	Additive noise. Extremely problematic to ASRs, particularly if speech over speech.
Other speakers	Background speakers, crowds/babble, crossed line.	Additive noise. Can influence ASR outcomes significantly if at a sufficient level. Next to impossible to remove post recording.
Interference/electrical	Mains hum, GSM (See Chapter 6).	Additive noise. Often evidenced through addition of harmonics (lateral plane). Can significantly influence ASR outcome.
Distortion	Signal exceeding capture resolution/clipping, microphone overloading or popping (plosive energy).	Often introduced at the recording stage, a compressor on the front end can partially mitigate. Anti-clipping tools (post) can also assist, though have marginal benefit with unknown impact on ASR outcome.
Digital/Analogue corruption (e.g. police interview cassettes).	Aliasing, glitches, age of media/proximity to magnetic field/heat (particularly analogue).	Additive noise. Aliasing, sometimes described as a ‘ghost mirror’ of the signal in the lateral plane. Jitter and shimmer in the analogue/tape domain (wow & flutter).
Signal loss	Drop out(s).	E.g. faulty equipment, broadband packet loss or incorrect speech detection.
Frequency limitation (Chapter 10)	Band pass filter or loss in frequencies, often due to the capture process or transmission/transcoding.	Frequency bandwidth often constrained deliberately to limit data (cost, efficiency).
Proximity	Drop in SNR (Signal to Noise Ratio).	Speaker(s) move. Recording device can often move. Double distance = $\frac{1}{4}$ speech energy (inverse square rule).
Transcoding (Codec) (Chapter 11)	Data loss/‘Moth holing’ caused by data compression.	Can introduce degradation. ‘Lossy’ transcoding likely to influence ASR outcome. Codec history may not always be known (e.g. uploaded material to the internet).
Bit depth/rate		Most commercial ASR systems operate at 16bit, 8kHz (sample rate).
AGC	Automatic Gain Control.	Mobile phone circuits mitigate against background noise not to ASR benefit.
Microphone response	Range, type, sensitivity.	Proximity, direction, shape (cardioid etc.).

With regards to intrinsic conditions and disfluency (Table 3.7) a study by Hughes, Woods and Foulkes (2016) on disfluency/filled pauses found benefit from extracting dynamic measurements (and nasals) from filled pauses to produce a discrimination system capable of an EER% of 4.08% and Cllr 0.12 (2016: p.126).

Intrinsic and extrinsic variables are often linked in a complex matrix of interdependencies (see 8.4 pertaining to SNR and the Lombard effect). In summary, intrinsic variability is unavoidable and can influence amplitude/sound pressure and pitch. Clearly a many variables cannot be influenced or controlled by the practitioner. From experience, control over the recording conditions would be rare, especially for both known and unknown speech samples.

It could be suggested that extrinsic variables introduced into the signal chain can be rectified through post recording processing, or audio enhancement treatments. This will be explored in the experiments pertaining to SNR. It is suggested that almost all audio enhancement (pre-ASR analysis) is likely to have relatively low efficacy unless pertaining to the controlled removal of known or predictive noise - e.g. reference cancelling (Alexander, Forth and Tunstall, 2012) or adaptive noise reduction (Künzel and Alexander, 2014). This is because, it is argued, that degradation is often caused by (irrecoverable) data loss. In other words, the complete obscuring of one audio signal by another in the same frequency domain is such that removal of one cannot reveal the other. A simple analogy of this might be a photograph of a person holding a bright flashlight pointed directly at the camera. Attempts to digitally remove the highest intensity light, from a photograph, would not reveal the image 'behind' the glare. Degradation of the speech signal from extrinsic variability is common and unavoidable and so, as users of ASR systems, it is therefore important to understand both the intrinsic variability of speech and the extent to which extrinsic (acoustic) variability influences ASR performance.

### **3.3 Audio Recording and The Signal Path**

This section provides a summary technical explanation as to audio capture (microphone), digital recording (sample rate, bit rate and depth), transcoding/data compression and frequency bandwidth.

#### **3.3.1 Audio Capture**

The microphone is the first point in the recording/transmission process and arguably one of the most important. There are many types of microphone. The carbon button, designed for early telephones was invented in the 1870s and is attributed to several early pioneers (Hughes, Berliner, Edison). Dynamic, or moving coil microphones gradually replaced these and they are arguably the most prevalent today. Capacitor microphones, popular in recording studios due to their quality, were pioneered by Wente in 1916 and the electret was invented by Sessler and Bell in 1962. In

conjunction with capacitor technology electret technology paved the way for modern MEMS microphones (Micro-Electro Mechanical Systems) which are commonly found in smartphones today.

Excluding obscure types, such as laser or array systems, microphones operate under similar principles. Acoustic energy, sound waves, hit a sensitive diaphragm which then vibrates. Movement from the vibrating diaphragm is transferred into analogue electrical energy using either the magnet and coil or capacitance/electret principle and passed to the analogue to digital convertor for digitisation. Increasingly the technology of microphone and convertor is combined in extremely small, low profile units (MEMS).

When considering speech recording it is important to note the complex interactions that occur between the speaker(s) and microphone. For example, an individual knowing that they are being recorded (or not) and their compliance with that process could influence speech output (e.g. whispering/shouting, conspiratorial tone, withdrawing/monosyllabic, turning away or obscuring the microphone, deliberately increasing distance etc.). The angle, position, distance, response and type of microphone will vary the range of speech frequencies captured too. The positioning and orientation of microphone and speaker(s) in a room could also influence the extent of reverberation and environmental noise (perhaps less applicable in telecommunications use where distance tends to be more stable). Speaker and/or microphone movement could produce a speech recording either too quiet or loud (distortion). Application of the wrong type of microphone could cause certain speech information to be absent from a recording or too poor in quality to process using an ASR system (Rose, 2013). For the purpose of the experiments conducted in this thesis a controlled corpus with known microphone conditions was used to minimise baseline variability.

### **3.3.2 Digital Recording and Sampling**

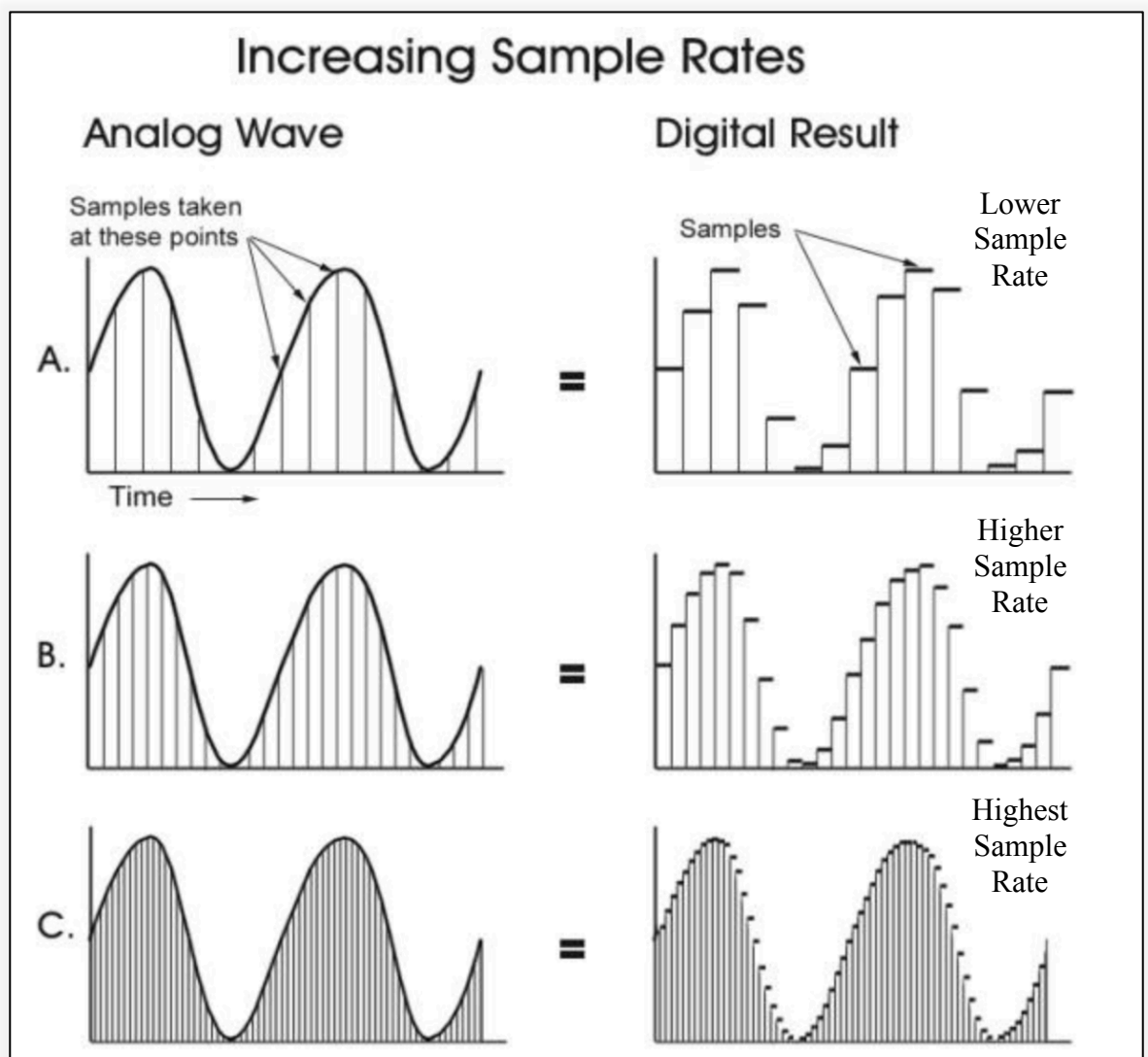
The electrical signal output from a microphone requires conversion or encoding into a digital format for transmission (e.g. telecommunications), digital recording or further computer processing such as editing and ASR processing.

The digitisation process (analogue to digital or A/D) is largely evolutionary but widely attributed to Alec Harley Reeves, a British telecommunications engineer. In 1938, Reeves designed and patented a pulse code modulation (PCM) coder/decoder (or codec) and provided the foundation for modern digital audio. Later during World War II, the Bell Telephone Laboratories developed a system of digital (PCM) transmission and reception with encryption. In 1943 this technology contributed towards a capability for the allied forces to provide encrypted communications between the UK and US (codename SIGSALY).



To summarise, the electrical signal output from the microphone is passed to an analogue to digital (A/D or ADC) convertor. The encoder works by producing a steady stream of numerical values at a given rate per second known as the sampling frequency. The values are then modified for each sub-second sample dependent on the incoming signal. The more samples per second (increase in sample rate) the more accurately the waveform is captured (Figure 3.9). However, a trade off in terms of quality is the quantity of data that the process can generate. This then has resource implications in terms of transmission and data storage with greater requirements for network bandwidth and memory.

**Figure 3.9:** Analogue to digital conversion and sample rate  
From: Dalemultimedia.com (+ annotation)

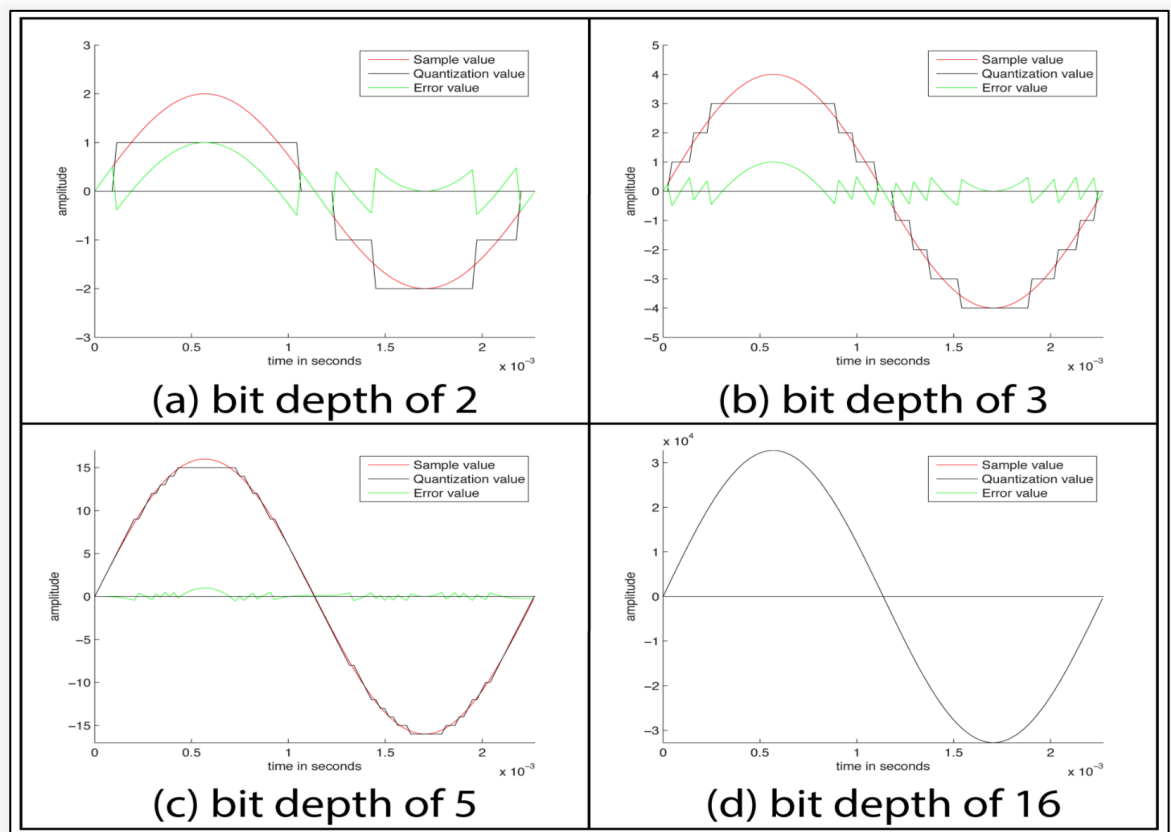


To listen to the recorded digital signal is then passed back through a digital to analogue (D/A or DAC) convertor. There it is converted to an electrical signal (speakers/headphones) and finally sound waves to the ear.

Harry Nyquist worked in the AT&T/Bell research and development laboratories from 1917 to 1954 and he and his team pioneered much of the research conducted on digital sampling and signal processing. This included research on optimum sample rates and the prevention of aliasing. Aliasing occurs when the analogue input wave contains frequency content beyond the range of that which can be digitally converted and represented. Nyquist stated that the sample rate should be twice the highest audio frequency to be digitised (Nyquist Rate). This prevents aliasing and best represents the incoming analogue signal in its entirety, in digital form.

Bit depth in the context of PCM is another important variable related to digital audio. Not to be confused with bitrate (number of bits transmitted or processed per second) bit depth refers to the quantisation level of the values in the vertical aspect of the converted waveform. Bit depth is also related to dynamic range which is the range of quantised values from the lowest, quietest, level audio signal that can be converted to the loudest. Higher bit depths (such as 16 and 24) equate to compact disc and 'studio quality' recordings with lower bit rates (e.g. 8 to 12) used extensively in telecommunications, largely due to higher bit depths equating to more data. Low bit depths can sound synthetic as greater quantisation of the input waveform produces a less continuous and more stepped digital waveform which can be audible. There is also a higher likelihood of error and crude/incorrect representation of the analogue waveform and this is illustrated in Figure 3.10.

**Figure 3.10:** Analogue to digital conversion and bit depth  
From: Digitalsoundandmusic.com



### 3.3.3 Transcoding (Codec)

It is necessary to first differentiate between encoding and transcoding, which are commonly confused. Encoding is to convert a signal from analogue into digital. Transcoding in the context of audio refers to the conversion of one digital format to another and the word codec is simply formed from the two words coder and decoder.

There are literally hundreds of different codecs available and each can have multiple settings which produces near infinite combinations. It may seem preferable for the audio community to have fewer codecs to provide consistency and limit variability. Nevertheless, there are valid reasons why a particular codec might be used in preference over another or regarded as unsuitable or redundant. Codecs can have different applications (e.g. optimised for speech and/or music) and are also continually evolving in terms of performance. If the digital output (codec) of one system is not compatible with the input of another system it may be discarded or updated. A cheaper codec (e.g. open source) may be preferred to one incurring a licensing fee. A more efficient codec that utilises less data may be required (i.e. cheaper for transmission and storage). Also, a high degree of audio quality may be deemed of lower importance in comparison to the more simplistic provision of speech intelligibility, particularly when attribution is known or not required. One example of this might be push-to-talk radio (PTTR) communication systems with limited users or call signs/self-identification where frequency bandwidth can be constrained. With PTTR or mobile 'phone devices the distance between microphone and speaker is also usually small and so a codec's settings can be set accordingly – with less requirement to mitigate against poor proximity and/or ambient noise.

### 3.3.4 Bit Rate

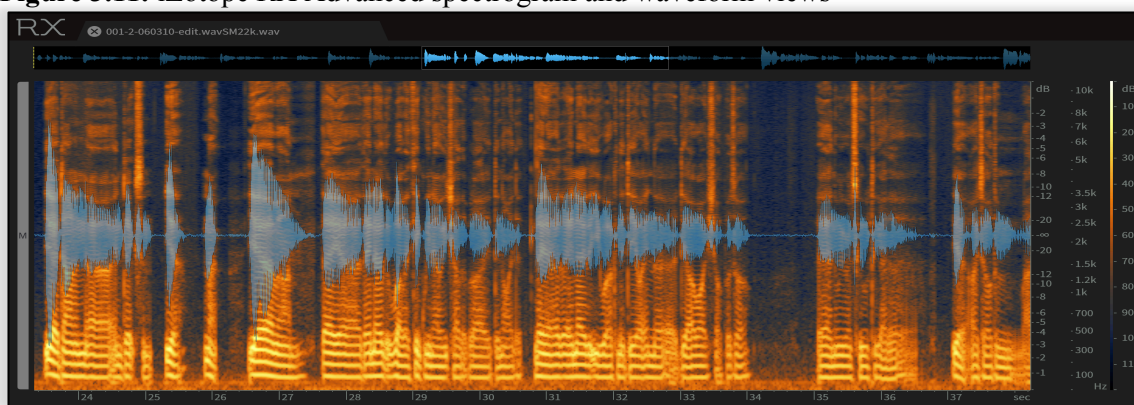
Bit rate refers to the number of kilobytes (data units) per seconds. In the context of audio recording and for pulse code modulated (PCM) sampled audio - bit rate is equal to the sample rate multiplied by the bit depth multiplied by the number of audio channels (i.e. x2 for stereo). To provide context the bit rate of a standard compact disc would be  $44.1\text{kHz} \times 16\text{bit} \times 2 = 1,411.2 \text{ kbps}$ . Bit rate is pertinent to transferring data, transcoding and audio quality and many codecs are considered 'lossy' as they effectively reduce digital information by compressing data. Some codecs may also have an option to utilise a variable bit rate (VBR) - e.g. MP3 - to adhere to adaptations in network transfer speeds, for example. In summary audio/speech quality (frequency bandwidth and/or digital rendering detail) is sacrificed when transcoding to a low bit rate.

### 3.3.5 Spectrogram Analysis

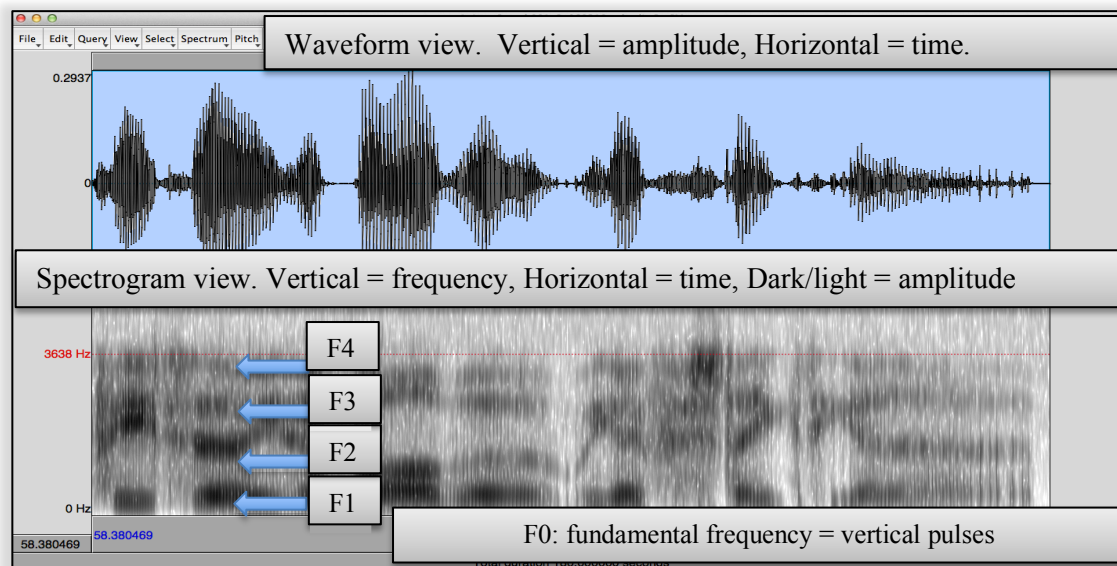
Speech can be viewed using spectrogram analysis as found in applications such as Izotope RX Advanced ([iZotope.com](http://iZotope.com)) and Praat ([fon.hum.uva.nl/praat](http://fon.hum.uva.nl/praat)) (see Figures 3.11 and 3.12). Spectrograms enable an audio analyst to examine frequency content (y axis) against time (x axis).

Figure 3.11 illustrates the (iZotope) spectrogram of a section of speech. The orange colour intensity denotes amplitude by frequency which, by default, scales from -120db to 0db (Figure 3.11). The horizontal orange lines shown in Fig 3.11 are the harmonics (multiples of the fundamental frequency). The harmonics are characteristics of the voice source (larynx), whereas the formants (= vocal tract resonances) are characteristics of the filter (vocal tract). The blue waveform displays summed amplitude (all frequencies) in the time domain. Izotope RX Advanced is commonly used for acoustic examination and includes powerful tools for altering audio in the frequency/time domain. By default, the frequency scale displays using Mel (see 3.4.3) but other scales can be selected (e.g. linear or logarithmic) which enables intricate acoustic examination for other types of audio events (e.g. music).

**Figure 3.11:** iZotope RX Advanced spectrogram and waveform views



**Figure 3.12:** Visual representations of speech in Praat, amplitude and spectrogram



Praat is predominantly used by speech analysts and phoneticians. The spectrograms differ in respect to the display and can be extensively configured. Praat defaults with Gaussian windowing (rather than Hann in iZotope) and the dynamic range of Praat defaults to 0db to 100db. Figure 3.12 from a section of DyViS speech. "...Peter, he's a barber, we go for steak together..." Formants (F1, F2, F3 and F4) are represented by dark horizontal lines on the spectrogram. F0, the fundamental frequency, is represented by the fine and feint vertical lines show the glottal pulses (spectrogram view only). Praat (Figure 3.12), with its adapted display settings, better represents formant data visually and is

therefore more suitable for auditory phonetic analysis. The main difference between the spectrogram shown in Fig 3.11 and Fig. 3.12 is that the former is a narrowband spectrogram whereas the latter is a wideband spectrogram. Only in wideband spectrograms can the formants be seen clearly. The variable that determines the difference between these two kinds of spectrograms is the length of the analysis window. In summary, spectrograms are essential for assessing the technical quality of speech recordings, analysing acoustic degradation and examining noise and speech in detail. Izotope RX Advanced and Praat were used extensively throughout the experiments conducted to analyse speech files.

### **3.3.6 Signal to Noise Ratio**

Signal to noise ratio (commonly referred to as SNR or S/N) is the strength of the desired signal in comparison to the unwanted - i.e. noise. SNR is measured in decibels (dB) with 0db at the equal ratio of signal and noise. Positive values are generally referred to as high SNR and negative values are referred to as low SNR with noise effectively beginning to obscure the desired signal. Measuring, or rather estimating, SNR is difficult since the ratio varies throughout the audio file (Beritelli et al., 2010). Additionally, there are different ways of representing the db output too, such as A-weighting to account for the frequency response of the ear (Fletcher and Munson, 1933). Debate endures as to the best way to correctly measure SNR although it is widely accepted that accurate estimation is a more preferable expression of the term than absolute measurement. The simplest way to estimate SNR is to divide the power of the signal by the power of the noise. There are several ways of measuring power including popular methods such as the root mean square (RMS), peak or loudest amplitude and the mean amplitude. It is also widely accepted that estimation of SNR becomes less accurate when noise is high. Martin (2001) developed an innovative new algorithm using a technique known as minimum statistic noise estimation. This was further progressed by Kim and Stern (2008) with Waveform Amplitude Distribution Analysis or WADA, generally agreed to be a more robust method of SNR estimation when noise is high. Kim and Stern achieved this through improving the discrimination of speech (over noise) and recognising that speech is predominately represented by a Gaussian distribution (see 3.4.5.1) in comparison to noise - which is generally not (excluding crowds/background speech or 'babble'). It is acknowledged that WADA SNR estimation could be prone to inaccuracies if measuring foreground speech against background babble which is also Gaussian in distribution. Nevertheless, as WADA SNR measurement is widely considered a robust method of SNR estimation and babble is not used in the experiments in this thesis, WADA is therefore applied in this thesis.

### **3.3.7 Reverberation**

Reverberation is the reflection of sound waves summing with the original signal. Direct sound is that which travels straight from source to listener (or microphone). Non-direct sound, for example room reverberation, consists of reflections of the direct sound from surfaces such as the walls, ceiling, windows and furnishings which return back to the listener or microphone after a small delay.

Dependent on the temperature of the air the approximate speed of sound is 343m/s (at 20°C) and the difference in time, between the direct sound and the non-direct sound (to the listener), is described as early sound. Reverberation is both frequency and amplitude dependent and is not limited to indoor spaces (e.g. mountain ranges). The length or duration of sound reflections (total reverberation) is measured as the time taken for the sound to diminish in amplitude by 60db once the sound source ends. This measurement is widely known as reverberation time 60 or RT60 (Schroeder, 1964). The time difference between direct and early sound arrival is often just tens of milliseconds, but (along with RT60) assists with providing the sense of space.

For the purposes of the experiments conducted in this thesis convolution reverberation is applied digitally. This process utilises a system of digitally capturing the impulse response (IR) of a 'real' reverberant space and then creating a mathematical model to reconstruct that space. This allows a highly controlled application of digital reverberation to the incoming signal. The detail of how IR is applied to controlled data is discussed in Chapter 9.

### **3.3.8 Frequency Bandwidth**

In the context of audio and speech recording, frequency bandwidth refers to the range of frequencies digitally captured. Frequency bandwidth is related to the sample frequency such that the highest possible recorded frequency is equal to half the sample rate. As discussed this is known as the Nyquist frequency and, by way of example, standard telephone communication is recorded at a sample rate of 8kHz, the highest speech frequency captured is therefore 4kHz and the frequency bandwidth 0-4kHz.

### **3.3.9 Channel**

The term channel (sometimes referred to as 'domain') is used throughout the experiments to define and differentiate the type of recording path, such as telephony or interview (room). Whilst outside the scope of the experiments conducted in this thesis, in a broader context the term channel could also be used to differentiate other types of speech recording - such as Voice over Internet Protocol (VoIP), push to talk radio communication, body borne recordings etc. Links between channel and ASR performance has been previously researched. For example, it is widely known that speech transmitted via the telephone channel is constrained both by frequency bandwidth and the GSM speech codec used to transmit/receive the signal. Both these channel specific variables are known to influence formant measurements and therefore ASR performance (Künzel, 2001; Besacier et al., 2000; Byrne and Foulkes, 2004). In addition, Hughes et al., (2019) stated that some speakers' formants are harder to measure/track than others and can vary across different channel types (2019: p.4).

Research into the effect of channel by Reynolds et al. (1995) also determined that the microphones used to capture telephone speech and the subsequent distortions produced influenced the accuracy of ASRs. To therefore avoid the conflation of additional variables it was determined that the effect of channel should be heavily constrained. Degraded baseline data from the same channel was therefore used in the experiments in preference to introducing cross channel variability.

## 3.4 Automatic Speaker Recognition Systems

This section provides additional technical explanation as to the terms relating to speaker verification systems. ASR systems require three types of speech data. These are defined as:

- i. A known sample(s) or speaker model(s) (one or multiple attributed speakers);
- ii. An unknown sample(s) or test audio file (one or multiple unknown speakers);
- iii. A data population or normative data/background model (multiple anonymous speakers).

### 3.4.1 Speech Detection

Setting aside data preparation (e.g. digital editing by the operator) an important process in an ASR system is to discriminate speech from non-speech. This is often referred to as speech detection (SD), speech/activity detection (SAD) or voice activity detection (VAD) attributed to Bennyassine et al., (1997)<sup>††</sup>. It is important to note that speech/non-speech discrimination can occur at different points of the file ingest chain (e.g. pre or post feature extraction).

On enrolment onto an ASR system both the speaker models and the test audio files often have speech detection applied to them prior to, or as part of, the feature extraction stage (3.4.3). In OWR iVocalise there are settings that pertain to VAD, i.e. detection of silence at the sub-second level and subsequent removal - with options available to the practitioner to turn it off completely, which will be discussed in later chapters.

Speech detection as a concept largely predates ASR system designs. Early algorithms were developed at around the same time as the early pattern matching systems but were mostly used to assist with locating sections of speech on multiple radio channels. An example of this was Dabbs and Schmidt (1972) application of speech detection in support of the NASA Apollo space missions. In their communications systems they had observed that the main power band of speech occurred in the 400Hz to 800Hz range and had a significantly higher signal to noise ratio (SNR) – i.e. F1 and low frequency F2. Dabbs and Schmidt used this information to devise an algorithm to successfully detect speech in radio signals which were reported to have an accuracy of approximately 90%. Speech research was arguably well funded throughout the 1970s by the US Department of Defence

---

<sup>††</sup> See also El-Maleh and Kabal (1997) for technical summary of different VAD approaches.

and algorithms progressed in complexity and accuracy. Later Saunders (1996) successfully used SD to discriminate between music and speech on the radio and significantly influenced modern algorithms - utilising a combination of amplitude, power band and the Gaussian distribution of speech to improve discrimination. VAD was further developed for variable-rate communications use by Sohn, Kim and Sung (1999). High quality speech detection is essential to accurate ASR system output and is very pertinent to the experiments conducted. In addition, speech detection can provide a difficult set of operating thresholds to balance. For example, if speech detection is set too aggressively then the ASR may lack sufficient speech information for successful comparison (inhibiting). If the speech detection thresholds are not set high enough then the ASR will attempt to apply modelling and comparison to non-speech sounds such as background speakers, noise or media such as TV or radio (contamination).

### **3.4.2 Diarisation/Speaker Separation**

Diarisation refers to the automated and semi-automated speaker segmentation processes most often applied to the test audio. Although manual editing and then speech detection (to remove silence) were used in the experiments in this thesis in preference to semi-automated diarisation which is generally regarded to be less accurate – it is nevertheless an important part of most ASR systems. Whilst somewhat outside the scope of this thesis - see Miro et al., 2012 for a review of diarisation research - diarisation is often confused with speech detection (which is also a process within diarisation) and so brief technical explanation follows.

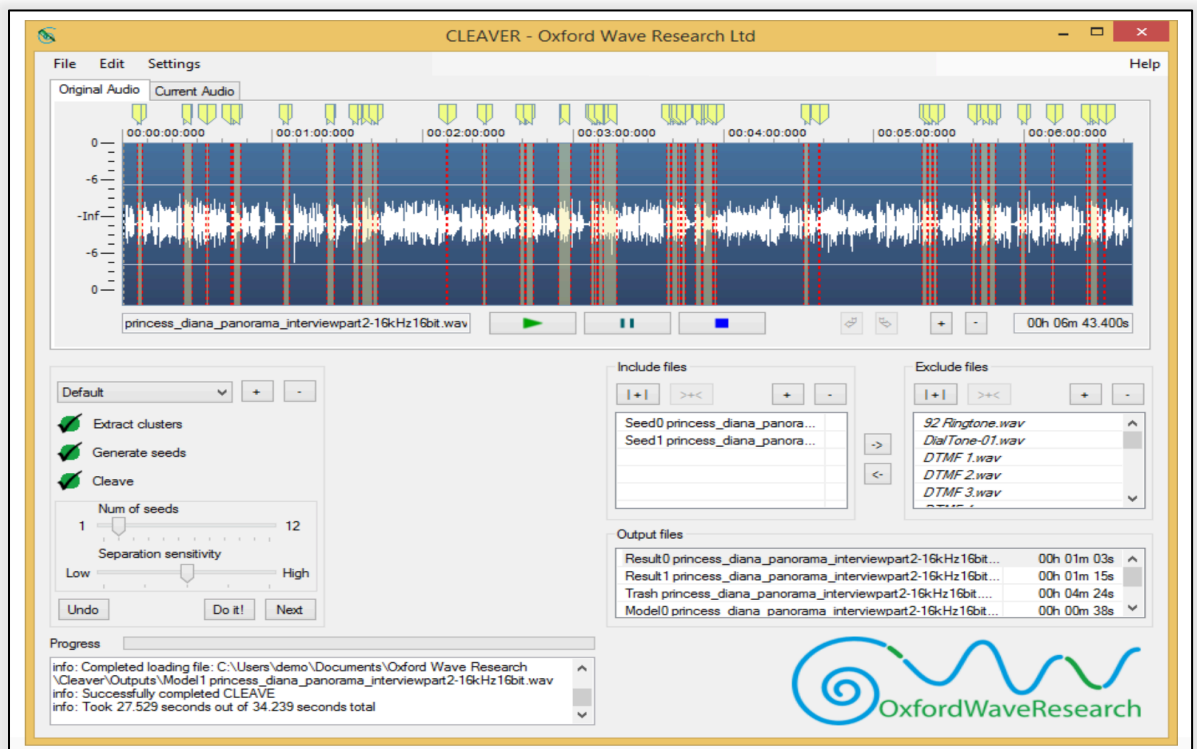
In summary speech is detected and separated from non-speech. Speech is then clustered or binned into multiple (or single) speech files dependent on broad similarities between speakers. These bins are then determined as speaker 1, 2, 3, undefined etc and concatenated ready for ASR processing. Many diarisation tools are command line operated. Diarisation can be a useful data preparation method of audio recordings for ASR system analysis. It can, for example, be applied at scale to batches of mono files (e.g. telephone conversations) to pre-process questioned speakers for ASR use. Diarisation is generally not recommended for application in creating known samples/speaker models. A higher degree of control and human interaction is important since the files are used for recognizing other speakers.

Tranter and Reynolds (2006) provided a detailed overview of diarisation, a process which automatically analyses an audio recording of multiple speakers (usually 2 or more) and attributes portions of speech to each speaker(s). A recent system called CLuster Estimation and Versatile Extraction of Regions or CLEAVER (Alexander and Forth, 2012) designed by Oxford Wave Research is an example of a diarisation application with an intuitive graphic interface. This provides the user with control of multiple settings including the sensitivity of speaker segmentation and elimination of non-speech sounds (an exclusion database). In Cleaver the user can also choose to



assist the process or automate ‘blind’ clustering to segment speakers accurately and in batches (Figure 3.13).

**Figure 3.13:** An example of modern, standalone, diarisation software (Clever by OWR)

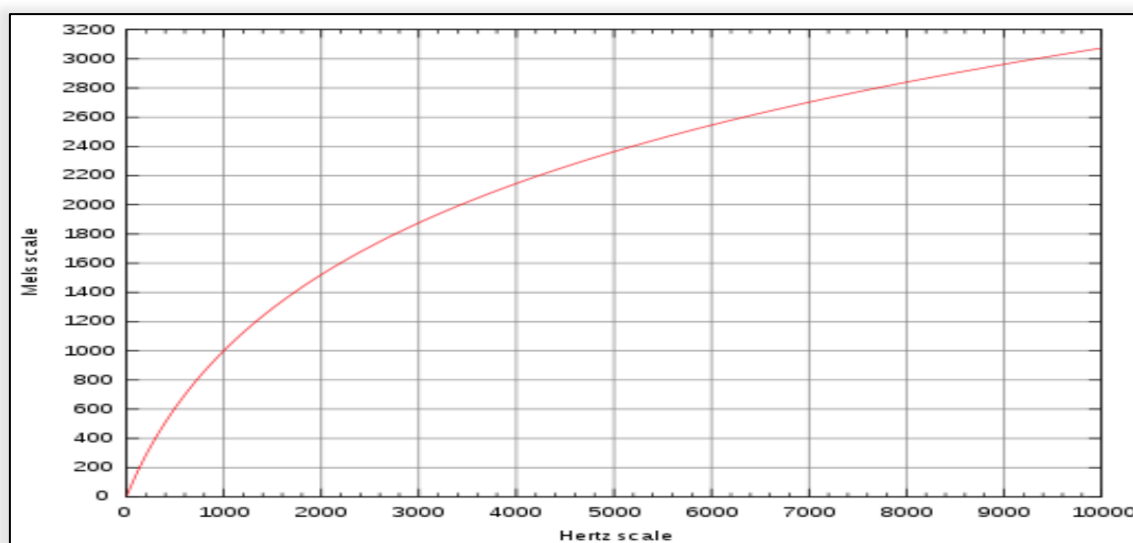


Small-scale experiments with different diarisation tools were conducted during the course of this thesis to determine if they could be utilised – particularly for the creation of speaker models. Broadly, they worked extremely well, but were found to be quite reliant on high divergence between speakers (i.e. how dissimilar they are). Accuracy also declined when the technical quality of the recordings was lower. When applied to DyViS data, cross-speaker contamination was therefore inevitable (high similarity) and unwanted truncation also occurred on degraded speech, especially on softer utterances (e.g. lower vocal effort/SNR). Settings were also difficult to define across multiple speakers, likely due to the variation in SNR. It is suggested that, whilst diarisation tools can bring benefit of speed and are useful in scalable systems for large scale processing – quality is compromised. Finally, the high quality and integrity of control data (validated speech samples/speaker models) is an extremely important aspect of a speaker recognition system. Diarisation tools were therefore rejected for the purpose of the experiments conducted. Manual editing (i.e. by hand) was found to provide much greater accuracy, less cross contamination and very low instances of speech sound truncation (loss of softer speech sounds). Manual editing also prevented any unwanted additional variability that could be introduced by the diarisation process itself.

### 3.4.3 Feature Extraction and Mel Frequency Cepstral Coefficients

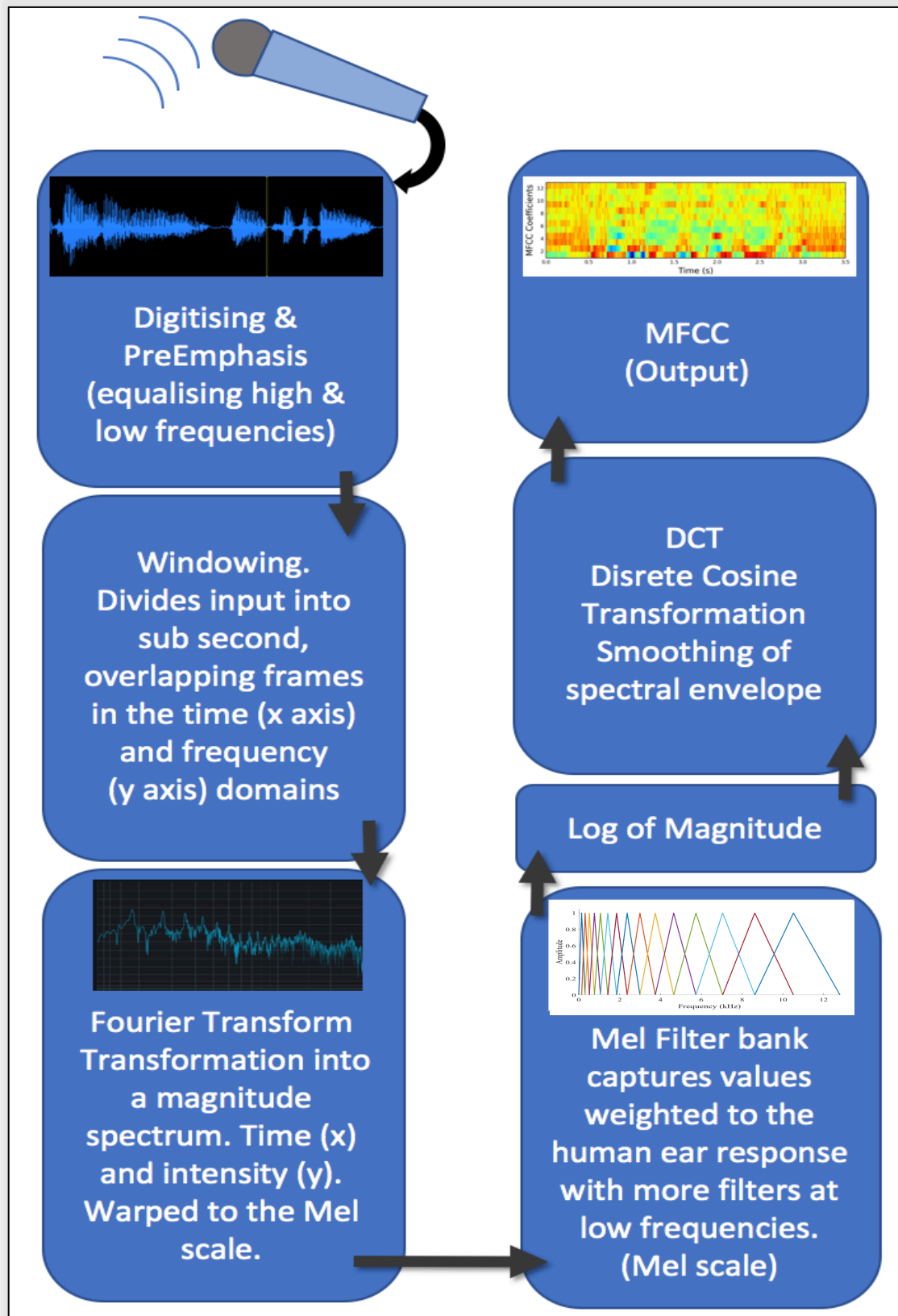
An important step in the ASR system is to extract data from speech to produce a statistical model. To complete this process the software performs a feature extraction. In computational terms the objective is to effectively represent the speech and the speaker in an efficient manner without using superfluous quantities of data. Mel-frequency cepstral coefficients (MFCCs) are considered the most common, reliable and proven way to represent vocal tract resonances. MFCCs have been in use since the late 1970s and are widely credited to Mermelstein (1976) and Mermelstein and Davis (1980), building on research completed by Bridle and Brown (1974). MFCCs, initially designed for speech recognition, are effectively numerical values simplifying measurements of the digital speech signal to enable data processing. To better explain MFCCs it is important to first clarify a few additional terms. Stevens, Volkman and Newman (1937) devised Mel as a frequency scale based on perceptual distances of pitch (the fundamental frequency). Mel is considered a pertinent scale to use for speech processing (e.g. speaker recognition and speech to text) as it correlates well to human hearing (Figure 3.14).

**Figure 3.14:** Mel and Hz scales  
From: [Deerishi.wordpress.com/tag/mfcc](http://Deerishi.wordpress.com/tag/mfcc)



Cepstrum refers to values calculated from the log of the power spectrum when the results are placed in the time domain as opposed to the frequency domain. Cepstrum takes the first four letters of the word spectrum and reverses them to reference that domain transformation. Coefficient simply refers to a numerical value, a variable. Note also that mathematical transformations are applied to the digital signal. These include Fourier transformation, which takes a time-based signal and applies filters to measure the intensity of individual frequencies and discrete cosine transformation (DCT) (Ahmed, 1972; 1991) which effectively smooths the detail from the spectrum and is also commonly used in compression algorithms. To perform a Mel Frequency Cepstral Coefficient (MFCC) extraction the following steps are taken (Figure 3.15).

**Figure 3.15:** Summary of the MFCC feature extraction process

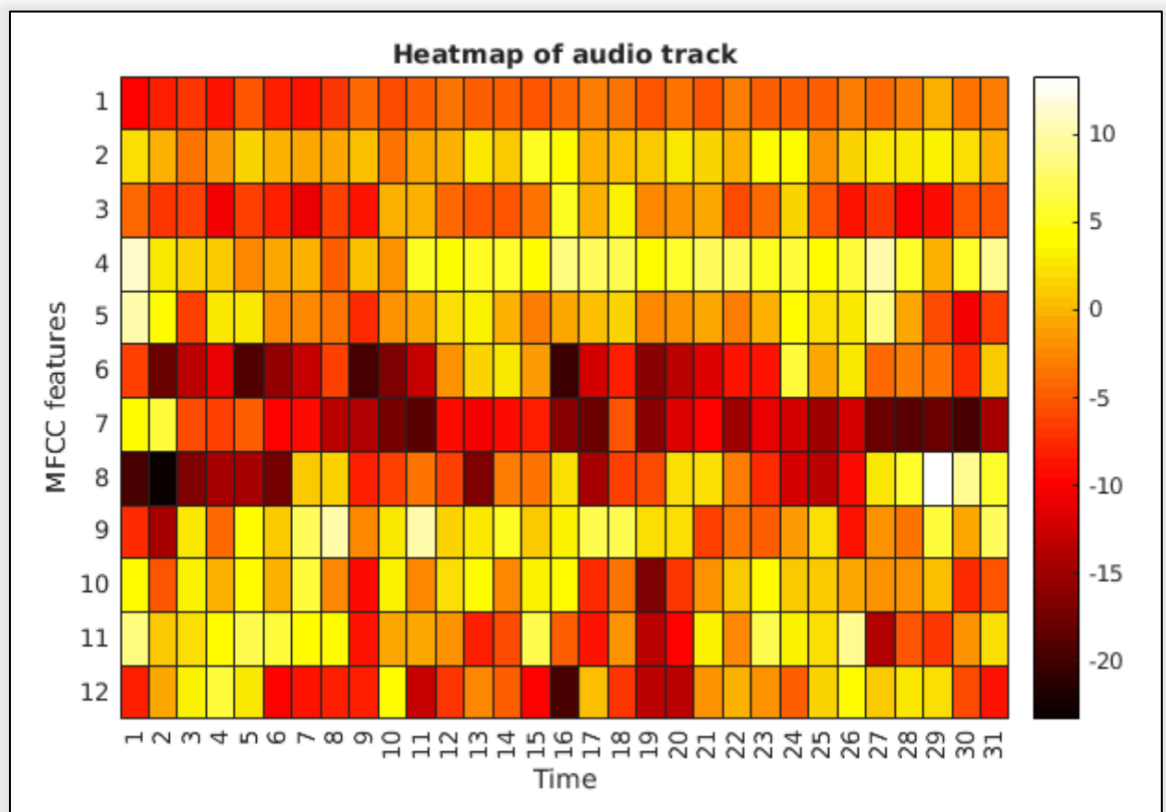


Mel filter bank and MFCC output images: aalto.fi.

Figure 3.15 was drawn from explanations from Beigi (2011: p.173), Furui (2001: p.253) and Fedila, Bengherabi and Amrouche (2018: p.16723).

MFCCs are often visually represented as heatmap grids as seen in Figure 3.16 (Lode, et al., 2018: p.5). A heatmap shows the values for each of the features (cepstral coefficients) in relation to time. In Figure 3.16, 12 features are extracted. The number of features in the MFCC extraction is a setting in Vocalise, as is the number of triangular filters. Both features and filters are later referred to in the preliminary tests (6.5.1). Note that one slight disadvantage of MFCCs is that the features are extracted in successive frames and are independent of each other over time. Algorithms are necessary to compensate for this and these effectively compute deltas and delta-deltas in the horizontal plain of the heatmap, i.e. differences, (usually from the mean) and longer-term change over frames. This, however, is not necessarily a disadvantage from a statistical perspective since movement from one frame to another (in the vertical axis) is independent and doesn't necessarily predict frame value (as shown in Fig. 3.16). MFCC's therefore enable a high degree of pertinent speech information to be passed to the feature extraction and statistical modeling stages.

**Figure 3.16:** Conceptual example of MFCC values/heatmap from Lode et al. (2018: p.5)



Note that there are many different methods of completing the feature extraction process. Previous research by Memon, Lech and He (2009) explored combining feature extraction methods and noticed a marginal uplift in EER% on YOHO and TIMIT data when fusing MFCC and IMFCC for their GMM-UBM ASR. Their study found TIMIT/GMM at 1.5% EER for MFCC and 1.8% EER for IMFCC with 1.4% EER for the fused system and YOHO/GMM 1.6% EER MFCC and 1.8 EER% IMFCC with 1.4% EER for the fused system. In their conclusion Boucheron and De Leon stated

the best feature extraction method was proven to be a fusion of the MFCC and IMFCC method (2008: p.4). Tirumalaa et al. (2017) recently provided a summary of the many different feature extraction methods that have been studied (to date).

### 3.4.4 Long Term Formant Distribution

Long term formants (LTFs) were presented by Nolan and Grigoras (2005) as a method of discriminating speakers using acoustic measurements of formants F1 and F2 from specific vowel utterances and diphthongs. In their research, also a case study, they extracted 4 acoustic measurements; of the vowel /ɪ/ (as in 'bit') and three diphthongs; /oʊ/ as in 'know', /aʊ/ as in 'mouth' and /ɔɪ/ as in 'boy'. These were used to successfully discriminate between speakers. This methodology provided the foundation of an alternative feature extraction technique and was applied by Becker, Jessen and Grigoras (2008) using semi-automatically extracted formant frequencies (long term formant values for F1, F2 and F3) on a controlled corpus containing 68 speakers. The group then statistically modelled the speakers (GMM) to achieve an equal error rate (EER) of 3%. Whilst performance was marginally less than MFCC feature extraction methods, this offered the potential for performance improvements in respect of cross channel analysis. This was later studied by Jessen and Becker (2010) and then further developed by Alexander, Forth and Jessen (2013) to provide long-term formant distribution (LTFD) analysis which was incorporated into an ASR (Vocalise). Vocalise LTFD was trialled in the preliminary tests and comparable EER% rates were observed, with results presented in chapter 6 (preliminary tests).

Extracting LTFD measurements in Vocalise is completed automatically by isolating information corresponding to the formants with a function call to the third-party program Praat ([fon.hum.uva.nl/praat/](http://fon.hum.uva.nl/praat/)). Praat then runs a script (extractVoiceandFormantsAA.praat). Extracting automated formant measurements in this way is not likely to be completely accurate (Harrison, 2013). Indeed, a more manual annotation technique was applied by Nolan and Grigoras (2005). Nevertheless, the Praat function returns estimated mean values for each formant (F1, F2, F3 and F4) for statistical modelling and comparison. In summary the Vocalise LTFD method completes the following<sup>§§</sup>:

- i. Pitch estimation occurs, across short segments applying auto-correlation which labels sections as voiced or unvoiced;
- ii. First 4 formants are extracted across short voiced segments using an LPC based method;
- iii. Formant estimations are returned to Vocalise;
- iv. Formant measurements are mean normalised or mean variance normalised and delta or delta-deltas are added (in Vocalise).

---

<sup>§§</sup> OWR, Alexander (2016) and Kelly et al. (2019).

It could be argued that the LTFD extraction process is very similar in nature to that of the MFCC process, capturing data from the audio signal in the spectral domain (acoustic). However, although MFCC extraction captures more information, it could be argued that less non-speech acoustic information is extracted in the LTFD process. A direct comparison study of LTFDs and MFCCs analysis was conducted in Gold, French and Harrison (2013) on DyViS data (task 2). Their study, using a bespoke Matlab system and vowel extraction method found LTFDs to be an effective method of discrimination (Cllr 0.9072 and EER 5.47%). In summary, although LTFDs offer the potential for improved cross-domain analysis (e.g. interview vs telephone) it was determined not to proceed with the use of LTFD methodology beyond the preliminary experiments in this thesis for two main reasons. The prevailing feature extraction method used by almost all ASRs is MFCC and also that LTFD performance was simply not as good as MFCC (GMM-UBM) when using formant estimation.

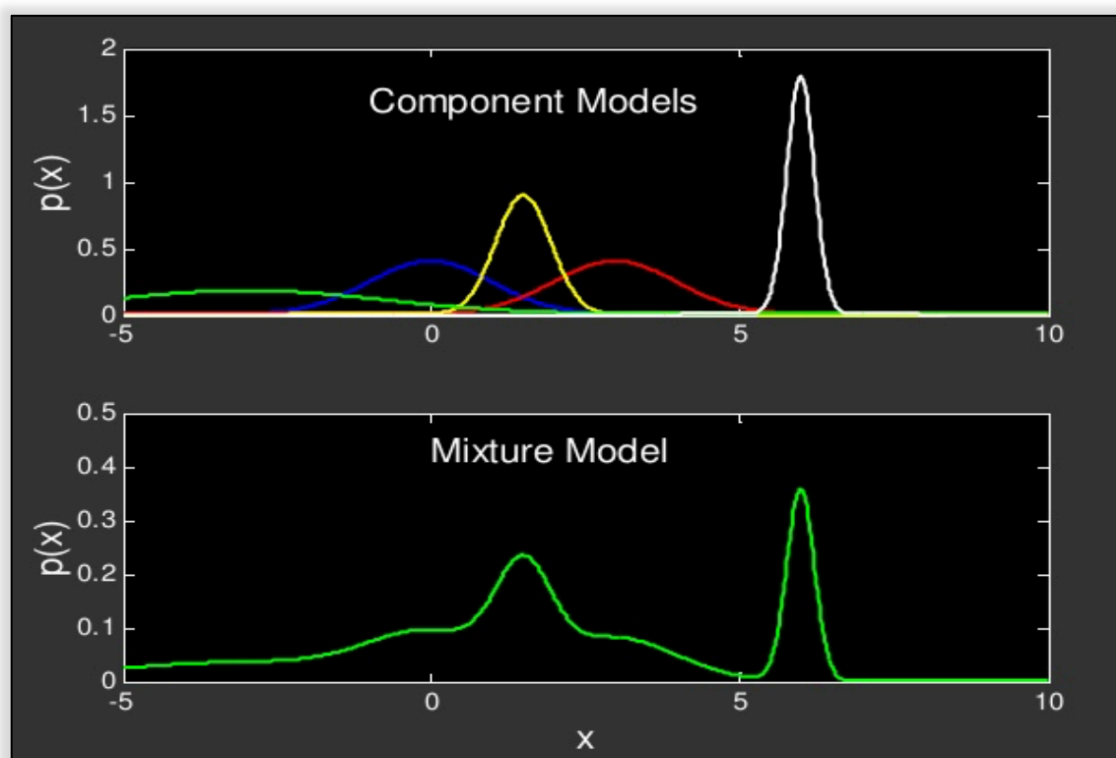
### **3.4.5 Statistical Modelling**

On completion of the MFCC (or LTFD) feature extraction process a statistical model is required to provide a summary representation of each speaker. An important point to note is that the statistical representation is therefore limited to the speech supplied and is not a comprehensive representation of the speaker. As discussed, the way in which statistical models are used has evolved from direct feature comparison (similar to pattern matching systems) to much more complex representations and the incorporation of comparative population or normative data. This section provides a brief technical explanation of the two main statistical modelling processes used in the experiments conducted (Vocalise/Gaussian mixture models and iVocalise/i-vector).

#### **3.4.5.1 Gaussian Mixture Models**

Gaussian mixture models (GMMs) are one of the most common ways of classifying and recognising complex patterns through the simplification and smoothing of data. Reynolds (1992; 1994) and Reynolds, Quatieri and Dunn (2000) are widely credited for applying Gaussian mixture models specifically to speaker verification systems. Reynolds (1992) discovered that GMMs were found to be particularly good for modelling statistical variation for speakers because they are able to represent a large class of sample distributions. They can therefore be used to model complex feature extractions from MFCC data with values relative to normative data. GMMs, via MFCCs, effectively provide statistical representations reflective of the vocal tract of a speaker (from the speech provided) relative to the universal background model (normative data). To provide further explanation, the term ‘Gaussian’ refers to the bell curve of a distribution (in this case from the MFCC output) and ‘mixture model’ applies to the layering of a number of those Gaussians components - see Figure 3.17 from Dulal (2014).

**Figure 3.17:** Illustration of 5 Gaussian components forming a GMM (Dulal, 2014)  
From: slideshare.net/dulalsaurab



In summary, Gaussian distributions are extracted and layered together to create single statistical models for each speaker relative to the mean GMM for the normative set of background speakers. Simply put, when an unknown speaker is then compared, their feature vectors are compared against the GMM of the known speaker to provide the numerator of the likelihood ratio and the feature vectors from the unknown speaker are compared against the UBM which provides the denominator. This is further illustrated in Enzinger (2015: p.52). Variances between the two speakers are measured (similarity) against the normative set (typicality) and this is what the likelihood ratio effectively equates to (see also 3.5.1).

As previously stated, Gaussian mixture models do not necessarily need to be created from MFCC feature extraction output and could be applied to other feature extraction methods. Also, the number of Gaussians that can be generated is configurable. In preliminary tests to examine this further (chapter 6) it was noted that the EER fluctuated marginally depending on the number of Gaussians selected – likely due to the difference in data detail/statistical density. However, diminishing performance was noted as the Gaussians increased – as also reported in Alexander, Forth and Jessen (2013). Whilst the reason for this is not entirely understood it is likely that this occurs as the statistical models become saturated with respect to the detail of the data extracted and inclusion of non-speech information occurs.

### 3.4.5.2 i-Vectors and Statistical Modelling

The introduction and development of i-vector systems, during the course of this thesis, heavily influenced the experiments completed.

Dehak et al. (2011) are widely credited with the application of identity vectors or i-vector statistical modelling to ASR systems. I-vector modelling was developed from joint factor analysis (JFA) by Kenny et al. (2006) - which, in summary, completed a statistical model of both speaker and channel separately.

I-vectors were increasingly integrated into ASR systems during the course of this thesis with a notable paradigm shift occurring from 2014 onwards as manufacturers of ASRs began to adopt the new method (e.g. Vocalise to iVocalise). At the time, i-vectors were viewed as potentially offering a performance advantage over traditional GMM methods through better statistical representation.

Following MFCC feature extraction, the i-vector method effectively enable both speaker and channel variables to be statistically represented in a much more detailed, multi-dimensional super vector space. This super vector space can include, for example, the number of MFCC coefficients (often including delta or delta-delta calculations) multiplied by the number of gaussians. The super vector can be as large as >30,000 dimensions and this undergoes a dimensionality reduction to create an i-vector (e.g. 400 dimensions). The development of i-vectors for speaker recognition was to provide additional density in statistical modelling and, therefore, more complex and specific representation of each speaker (from the speech provided) than GMM-UBM.

To provide a brief explanation, vectors are points in space that have direction and magnitude. I-vectors are therefore effectively multiple positions (vectors) in multi-dimensional space that can represent a highly complex statistical speaker model within a probabilistic space. The complete space, created from normative data, is called the ‘total variability’ or ‘total variability matrix/space’ TV(M) or TVS. This compact (i)vector is effectively a probabilistic factor analysis (Baum-Welch algorithm) of the GMM-UBM models created from the entire training set (normative data). Similar to a universal background model, in standard GMM-UBM only systems, i-vector ASR systems generally require a much larger combined normative dataset (population of speakers) independent to the comparison files.

A common method for comparing i-vectors is commonly referred to as ‘probabilistic linear discrimination analysis’ or PLDA. PLDA is essentially a comparison methodology, rather than a reference to the entire normative data set – although the term ‘PLDA session’ is often applied as such. Prince and Elder’s initial study focused on facial recognition and improving discrimination under poor lighting conditions or when subjects used different poses and expressions. PLDA



develops earlier methods such as earlier Linear Discrimination Analysis (LDA) by Fukunaga (1990) and McLachlan (1992) - which is also utilised in iVocalise, prior to the PLDA stage. Essentially, LDA then PLDA combine to further assist with maximising the between speaker variability whilst minimising the within speaker variability by creating a more discriminative space (with LDA reducing session variability).

An i-vector speaker model is trained from normative data (i.e. an i-vector extractor) which is, in turn, trained from MFCC's from a large set of trained recordings. The architecture of the iVocalise i-vector system is arranged as follows: UBM, TV, LDA and PLDA. The combination of this approach provides greater discrimination (than GMM-UBM alone) and prevents, to some extent, contamination of non-speaker information which might influence results (i.e. channel and noise).

For the purpose of this thesis and to constrain variables, since different ASRs use slightly different algorithms, population models and settings - all research experiments are conducted on either the Vocalise ASR (GMM-UBM) and/or iVocalise (i-vector/UBM, TV, LDA+PLDA) systems Alexander, Forth and Jessen (2013), Alexander et al. (2014) and Kelly et al. (2019a). Full specifications are provided in Appendix G.

### **3.4.6 Speaker Model or Voiceprint?**

In the context of the experiments conducted, a speaker model is a computer file that contains the statistical summary information of a speaker's vocal tract as extracted from an audio file containing speech from the speaker. Manufacturers often refer to these as voiceprints, which is misleading and application of the term 'voiceprint' can cause consternation amongst the forensic speech and analytical communities, for two main reasons. Firstly, the term is associated with the early widely discredited speaker identification technique that involved holistic and impressionistic, i.e. non-analytic, comparison of spectrograms (Kersta, 1962). Second, the term voiceprint could imply, through connotation, that ASR output is similar to fingerprinting and more conclusive than it is. As discussed, speech is highly variable and is neither a unique, fixed pattern mode of identification nor a 'true' biometric measurement in the sense that no direct physiological traits are measured, only vibrations in the airwaves. In addition, the speaker model is generated using only a very small example of speech from an (often brief) audio recording and acoustic measurements refer only to that which can be extracted from sound waves. Nevertheless, speech does carry substantial biometric information within it and individuals can be distinguished to a large extent on the basis of their speech patterns. So, to summarise, the term speaker model (SM) is broadly preferred to 'voiceprint' and is therefore applied in this thesis.

### 3.4.7 Normative Data/Background Population

Modern ASRs use a large dataset to statistically represent the population. This is called the normative dataset, the universal background model (UBM) or, in an i-vector ASR system, the universal background model, total variability matrix, linear discrimination analysis and probabilistic linear discrimination analysis (UBM, TV(M)/LDA+PLDA).

Normative data is required to establish statistical context by providing mean values and to inform the ASR system as to what speech is. The comparison of known sample and unknown sample provides statistical distance data on similarity whilst the population data provides data on typicality. It is generated from a large quantity of speech files, usually hundreds or thousands, from different speakers. Ideally, audio files are neither the questioned audio (unknown speaker) nor the speaker model (known speaker). Often the normative dataset consists of multiple commercially available corpora and it is relatively hidden from the ASR user - although some systems do provide options to create your own normative sets (Vocalise and iVocalise).

Rose (2013) recommended that, as forensic speech analysts, we should be prepared to obtain normative data for each case, although a debate regarding the selection of normative data for ASR speaker comparison has endured. It is argued that selecting population data requires time, patience, a lot of data to select from and much consideration to accurately and evenly capture the variability in speech (including aspects such as language, dialect, gender, duration, channel and recording conditions) to reflect a population relevant to the comparison(s) conducted.

In the context of GMM-UBM ASR comparison where  $H_0$  = the 'same speaker' hypothesis and  $H_1$  = the 'different speaker' hypothesis - Reynolds, Quatieri and Dunn (2000) stated that:

'...while the model for  $H_0$  is well defined, ( $H_1$ ) is less well defined since it potentially must represent the entire space of possible alternatives...' (Reynolds, Quatieri and Dunn, 2000: p.22)

'...There is no objective measure to determine the right number of speakers or amount of speech to use in training a UBM'. (Reynolds, Quatieri and Dunn, 2000: p.25).

Others suggest that whilst the size of the normative data set is important, it may not be as significant for ASR application as might be first thought and that diminishing returns of EER% performance are evident as saturation is reached through data quantity. Hasan and Hansen demonstrated that, by carefully selecting a diverse set of UBM speakers, the baseline system (GMM-UBM) performance could be retained using less than 30% of the original UBM speakers (Hasan and Hansen, 2011: p.1830) - although it is argued that this would require retesting with respect to a modern i-vector system.

ASR system designers approach the normative data problem in different ways with some manufacturers expecting full trust in a default dataset, which remains effectively hidden and unalterable by the user. It could be argued that this provides benefit through system commonality (repeatability and reproducibility) and reliability of output. Nevertheless, it can exacerbate the frustration that without a good understanding of the content of the normative dataset - true typicality cannot be measured. It also cannot be assumed that the dataset is reflective of the comparative samples. In the research domain meta-data is available (for research corpora) to assist with informing the normative selection and ensuring data relevance. For an investigative analyst/forensic practitioner, building a bespoke normative set can seem a sizeable requirement – given the unknown variables in the questioned sample. It could also prove a distraction from the case at hand and has resource implications with regards the time it can take to source appropriate speech files and test the dataset. In addition, issues could arise regarding audio laboratories validating results across multiple systems. From experience, it has been noted that ASR operators feel a strong inclination to use a default normative data set at the risk of depending on the manufacturer to determine how relevant (or not) the selection of that data is to the comparison.

Whilst not a key objective of this thesis, the testing of different sizes and types of normative datasets formed a small part of the preliminary experiments completed in this thesis and for seeking to mitigate against acoustic variability (see Chapter 9). For those purposes significant care was taken when selecting or adapting the normative dataset. It is clearly indicated when the UBM is changed or adapted and care was taken to prevent the conflation of variables.

## **3.5 Automatic Speaker Recognition Output and Performance Measurement**

As discussed, speaker comparison relies on measurements taken from sound vibrations in the air. The infinite degree of intrinsic and extrinsic variability has also been discussed and so measurements cannot be taken as absolute. Despite this, film<sup>\*\*\*</sup> and television often confuse fact and fiction and use simplistic shorthand terms to drive a narrative - such as a speaker ‘match’ or a ‘hit’. These terms often surface in the analysis community and should be discouraged. This section provides an explanation of the terms pertaining to ASR output that are more widely accepted.

### **3.5.1 Likelihood Ratio and Bayes’ Theorem**

Aitken and Stoney (1991: pp.20-21) suggested that there is a requirement for the output of a comparative process such as speaker comparison to fulfil the following criteria.

---

<sup>\*\*\*</sup> The film 2001 A Space Odyssey (1968) makes an early reference to voiceprint identification.

- i. To assess the strength of scientific evidence it is necessary to consider (at least) two explanations for its occurrence;
- ii. The evidence is evaluated by assessing its probability under each of the competing explanations.
- iii. The strength of the evidence in relation to one of the explanations is the probability of the evidence given that explanation, divided by the probability of the evidence given the alternative explanation.

Given the above criteria, Bayes' theorem for calculating probability was recommended for application to scientific disciplines (e.g. Aitken and Stoney, 1991; Evett, 1998). This concept was further progressed by Drygajlo, Meuwly and Alexander (2003) and Drygajlo et al. for ENFSI (2015) for specific application in the context of speaker comparison and ASR systems. It was also further explored in a study by Gonzalez-Rodriguez et al. (2004; 2006) and many of their research recommendations for compensating for the lack of data to improve ASR LR output estimation have been integrated into modern systems – e.g. Zhang and Tang (2018).

The ENFSI guidelines (2015: p.4 and reproduced in 3.18) also promotes Bayes and further defines what should be the province of the court or the expert in the context of speaker comparison. Note that prior odds information is additional data, which can be derived from sources not necessarily pertaining to speech or audio (2015: p.5).

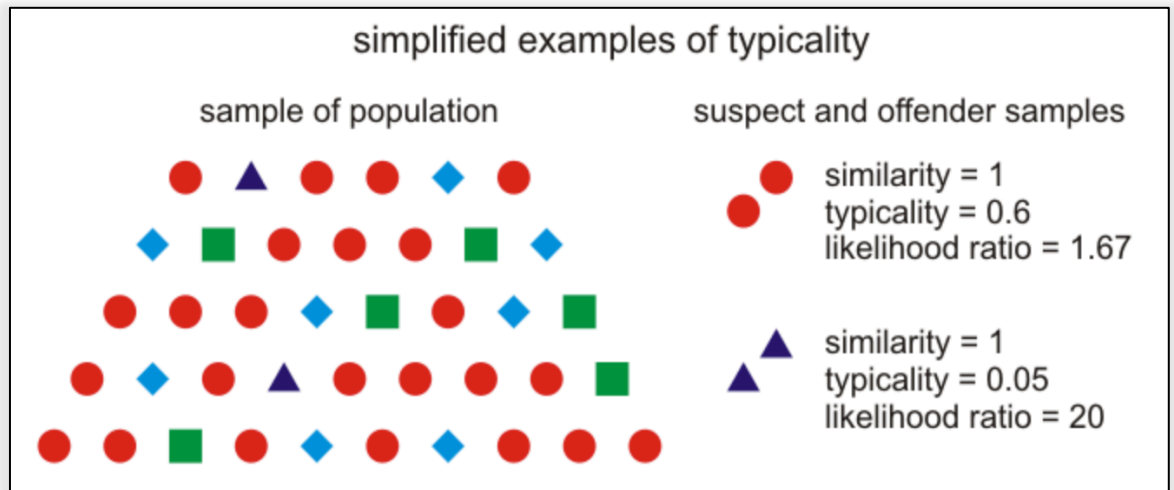
**Figure 3.18:** Bayes' theorem from Drygajlo et al./ENFSI (2015).

<p>posterior knowledge</p> $\frac{P(H_0   E)}{P(H_1   E)}$ <p><i>posterior odds</i> (province of the court)</p>	<p>=</p>	<p>new data</p> $\frac{p(E   H_0)}{p(E   H_1)}$ <p><i>likelihood ratio</i> (province of the expert)</p>	<p>×</p>	<p>prior knowledge</p> $\frac{P(H_0)}{P(H_1)}$ <p><i>prior odds</i> (province of the court)</p>
---	----------	---	----------	---

Likelihood ratio (LR) output from ASRs is the statistical probability of supporting either the same speaker hypothesis (H0) or a different speaker hypothesis (H1) and is effectively calculated from similarity divided by typicality. Note that some ASR systems (e.g. OWR Vocalise) presents output in terms of the Log of the LR or LLR.

Figure 3.19 from Morrison (2009) provides a further concise and useful graphic explanation of how LR is calculated.

**Figure 3.19:** Typicality, similarity and calculation of LR, from Morrison (2009)  
 From: [acoustics.org/pressroom/httpdocs/157th/morrison.html](http://acoustics.org/pressroom/httpdocs/157th/morrison.html)



Further studies pertaining to LR calculations are recommended in Aitken and Stoney (1991), Evett (1998), Hughes (2014) and Gold (2014).

Morrison, Ochoa and Thiruvaran (2012) proposed that the LR framework would be more accurate if the population database better supported the defence hypothesis (2012: p.62). They argue that speaker comparisons submitted, by the Police for example, were more likely to contain speakers which sound similar (and therefore generate a same speaker hypothesis) than to generate different speaker hypotheses. Their recommendation is that the selection of background and test data (e.g. channel and speaking style) is selected by a lay listener panel and put to a database. Their experiments showed benefit in an MFCC GMM-UBM system (over a randomly generated database) (2012: p.75). Whilst it is suggested that the argument to better support the defence hypothesis is sound, the approach could be difficult to implement with respect to time and resources given the permutations of channel and speaking style. In addition, it could be prone to errors pertaining to lay listener assessment. During the course of completing the experiments in this thesis it was noted that the specificity of the normative data was more important for GMM-UBM ASR system (performance) than for an i-vector system, where the requirement for normative data size was simply greater. It was also noted that the variation of LR or LLR output across ASR systems/normative sets could undermine the confidence of results. This problem was recently studied by Solewicz, Jessen and Van Der Vloed (2017) who applied a new method of score calibration to reduce the diversity in LLR output across 5 different ASR systems without the need for additional data.

Quantifying typicality forensically is ambitious due to the size of the data requirement (population). Research by Morrison and Enzinger (2018) examines the importance of incorporating information pertaining to the relevant population when calculating typicality for forensic application. They state:

‘Scores which are purely measures of similarity are not appropriate for calculating forensically interpretable likelihood ratios. In addition to taking account of similarity between the questioned-origin specimen and the known-origin sample, scores must also take account of the typicality of the questioned-origin specimen with respect to a sample of the relevant population specified by the defence hypothesis.’ Morrison and Enzinger (2018: p.1).

In summary, whilst prior and posterior odds are an important aspect of Bayes’ theorem, for the experiments conducted in this thesis the DyViS corpora used throughout provided statistically simplified data. For each of the 100 speakers every speaker model had a known test audio file (or multiple thereof). Finally, it should be noted that likelihood ratios (LR) values are not particularly easy to understand by the courts/lay-person. For other forensic disciplines verbal equivalence/interpretation of a numerical LR output is often offered by an expert to assist in understanding the strength of support for H0/H1. This is further discussed in 3.5.2.

### 3.5.2 Verbal equivalence scales

Verbal equivalence scales were proposed for forensic application by Champod and Evett (2001) and applied to speaker comparison by Rose (2002). The purpose was to design a scale to convert a numerical likelihood ratio (or log likelihood) score into an expression that a non-skilled person could better understand. The Table below is from Rose (2002: p.61).

**Table 3.20:** Verbal equivalence scale from Rose (2002: p.61)

Likelihood ratio	Proposed verbal equivalent	
>10 000	Very strong evidence to support . . .	
1000 to 10 000	Strong evidence to support . . .	
100 to 1000	Moderately strong evidence to support . . .	
10 to 100	Moderate evidence to support . . .	
1 to 10	Limited evidence to support . . .	
1 to 0.1	Limited evidence against . . .	Prosecution hypothesis
0.1 to 0.01	Moderate evidence against . . .	
0.01 to 0.001	Moderately strong evidence against . . .	
0.001 to 0.0001	Strong evidence against . . .	
<0.0001	Very strong evidence against . . .	

A more up to date and comprehensive guidance, Table 3.21, is also provided by ENFSI (2015) for evaluative reporting in forensic science and shows the evolution in phrasing in comparison to Table 3.20.

**Table 3.21:** ENFSI Verbal equivalence scale (ENFSI 2015: p.17)

<b>Values* of likelihood ratio</b>	<b>Verbal equivalent (two options of phrasing are suggested)</b>
1	The forensic findings do not support one proposition over the other. The forensic findings provide no assistance in addressing the issue.
2 - 10	The forensic findings provide weak support** for the first proposition relative to the alternative. The forensic findings are slightly more probable given one proposition relative to the other.
10 - 100	...provide moderate support for the first proposition rather than the alternative ...are more probable given...proposition...than proposition...
100 - 1000	...provide moderately strong support for the first proposition rather than the alternative ...are appreciably more probable given... proposition...than proposition...
1000 - 10,000	...provide strong support for the first proposition rather than the alternative ...are much more probable given... proposition...than proposition...
10,000 - 1,000,000	...provide very strong support for the first proposition rather than the alternative ...are far more probable given... proposition...than proposition...
1,000,000 and above	...provide extremely strong support for the first proposition rather than the alternative ...are exceedingly more probable given... proposition...than proposition...
<p>* Likelihood ratios corresponding to the inverse (1/X) of these values (X) will express the degree of support for the specified alternative compared to the first proposition.  **Forensic practitioners or their reports should avoid conveying the impression that a statement of the kind: "the forensic findings provide weak support for the first proposition compared to the alternative" is meaning that the findings provide (strong) support for the stated alternative. It just means that the findings are up to 10 times more probable if the first proposition is true than if the stated alternative is true. This is also the reason why the alternative should be explicitly stated. In cases where the reader could be misled as described above, forensic practitioners shall add additional comments.</p>	

Debate surrounds the use of verbal equivalence scales with one argument suggesting that the perception of verbal description can vary per individual (practitioner). This was researched by Mullen, Spence, Moxey and Jamieson (2014) and later by Marquis et al. (2016).

‘...results show that there are serious misunderstandings of the verbal scale. It does not achieve the purpose for which it was created. The terms used are unlikely to be understood properly by lay people and it would appear that they are actually misunderstood.’

Mullen, Spence, Moxey and Jamieson (2014: p.154).

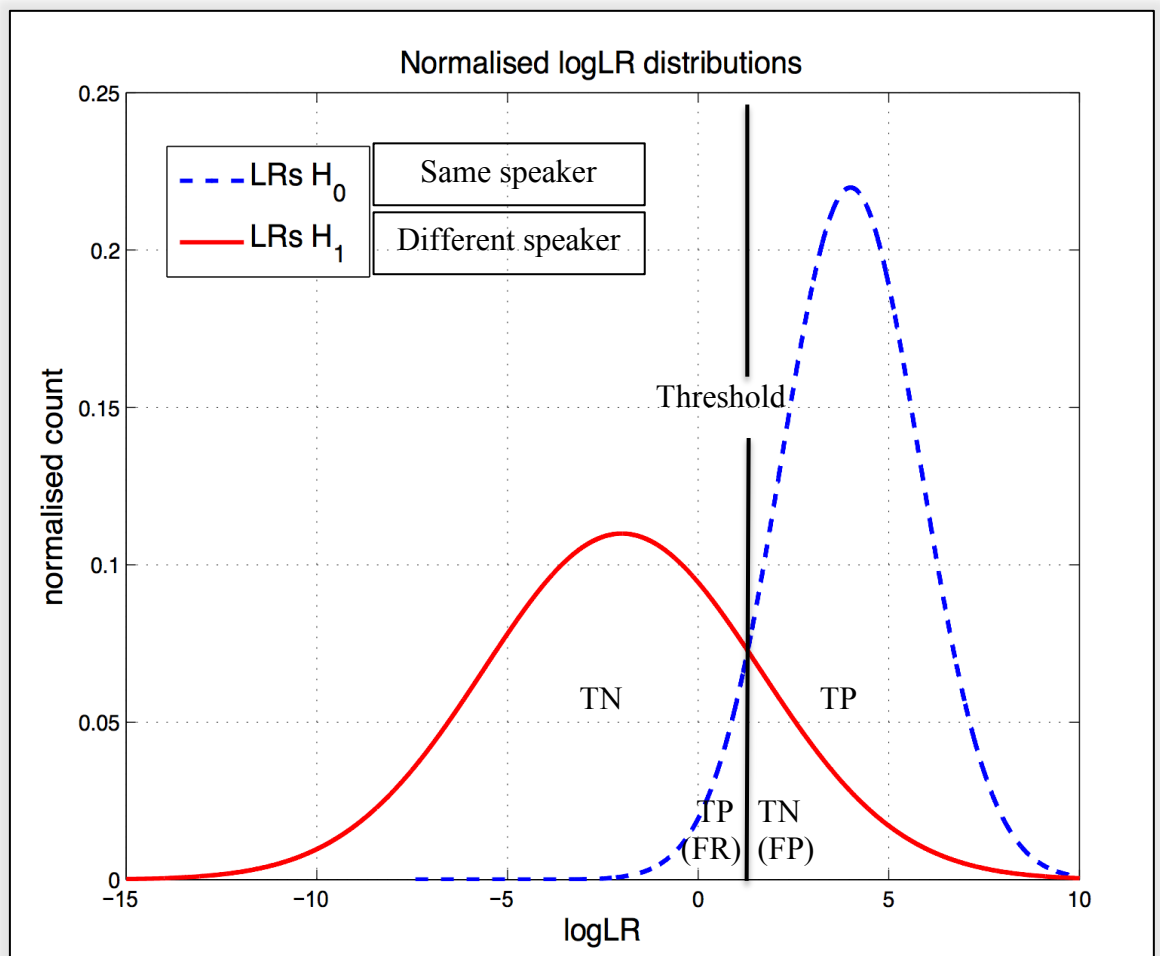
A second issue surrounds any simplification of the verbal equivalence table if log likelihood ratio conversions are applied (Vocalise). This creates the potential for cliff edge results with small margins to transition between descriptions (i.e. small numerical value differentiates between ‘strongly supports’ to ‘near certainty’). These concerns are countered by many in the community (e.g. Eriksson, 2012) who support a verbal scale which provides greater simplification. Eriksson (2012: p.60) also references similar verbal scales used by Swedish, Finnish, French and German law enforcement and states that output consistency consensus could be better reached between experts across LEA (repeatability and reproducibility). Nevertheless, the application of these types of verbal scales, including the use of phrases such as ‘near certainty’, are further discussed in chapter 12 in reference to the outcome from the experiments conducted.

### 3.5.3 Likelihood Ratio and Log Likelihood Ratio

#### Plots

When measuring ASR performance LR plots can be applied to illustrate the distribution of scores for multiple comparisons – some of which are same speaker ( $H_0$ ), some different ( $H_1$ ), Figure 3.22.

**Figure 3.22:** LR Plot example from ENFSI standards (2015: p.19) + additional annotation





Assuming a controlled corpus is used where the outcome is known there are four classes of results in the context of an ASR system and a sensible threshold can be identified which balances outcomes dependent on preference (Figure 3.22).

- i. True Positive (TP): the ASR has correctly verified the speaker (blue dotted line to the right of the threshold mark).
- ii. True Negative (TN): the ASR has correctly rejected the speaker (red line to the left of the threshold mark).
- iii. False Positive (FP): the ASR has incorrectly verified the speaker (red line to the right of the threshold mark under the blue dotted line) i.e. high LR score(s) for the incorrect speaker.
- iv. False Reject (FR): the ASR has incorrectly rejected the speaker (blue dotted line, to the left of the threshold mark under the red line) i.e. low LR score(s) for the correct speaker.

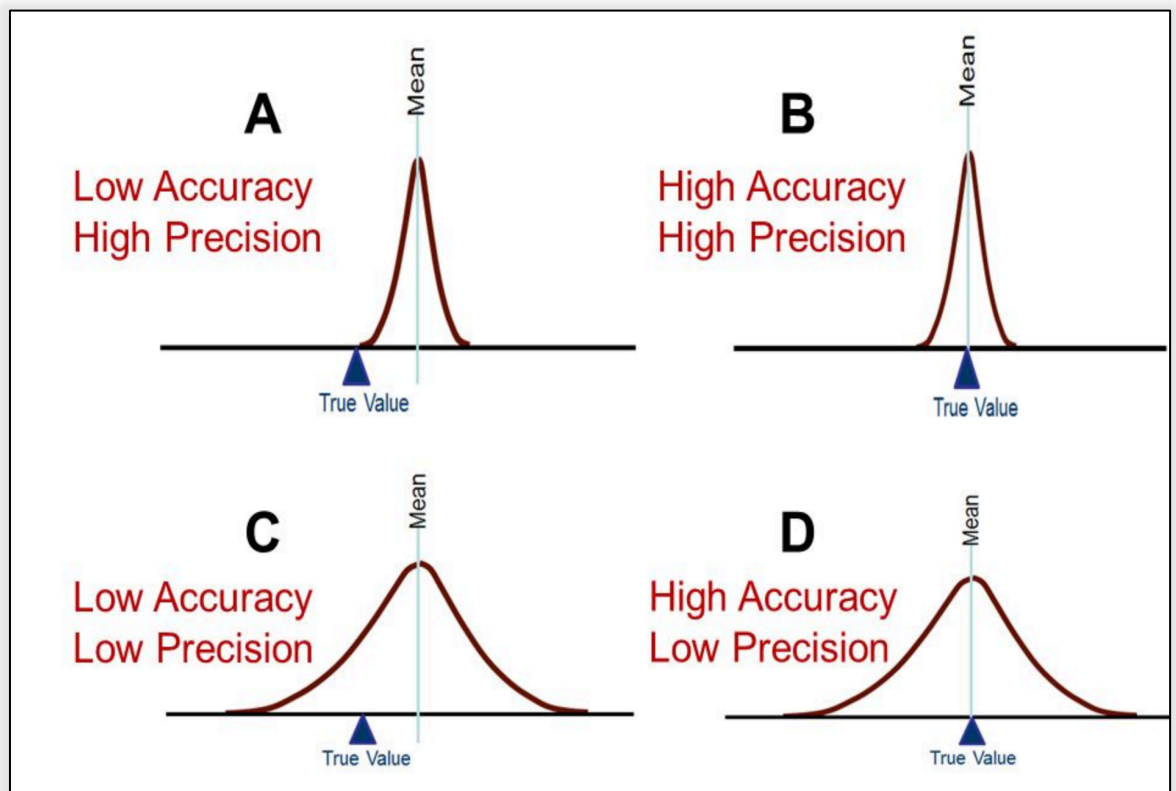
TP, TN, FP and FR terms effectively relate to same speaker distribution and different speaker distribution in relation to the threshold. The amount of separation between the same speaker and different speaker bell-curves is significant, with less overlap indicating better system performance and lower confusability. Greater separation between same speaker and different speaker distributions also provides the opportunity for clearer threshold setting. Conversely the closer the two distribution curves are the more likely it is that the system will provide incorrect output (FP, FR) with the setting of threshold values harder to determine with greater overlapping values. Note also that the score distribution should ideally provide a bell-curve with a narrow base for the same speaker results, again reflecting better overall system performance (less standard deviation). LR plots formed a key part of analysing the ASR output from the experiments in this thesis. In the experiments presented FPs and FRs are represented as a percentage of total outcomes and termed FAR (false accept rate) and FRR (false reject rate). To examine the trade-off between FP and FR, DET curves (and scatter plots) are also used in the field of speaker verification (Martin et al., 1997). When FAR and FRR are plotted the graphed lines intersect at a point to enable the calculation of equal error rate (EER%). Note that for some of the experiments conducted in this thesis the EER% is not quite zero despite the absence of FP and FR values. This is because EER% refers to the measurement of an area (under a curve) rather than a finite point. Whilst EER% as a performance measurement is not preferred by all forensic scientists (e.g. see cost of likelihood ratio (Cllr) 3.5.5) it is nevertheless widely used as a common way of comparing performance both within ASR systems and across ASR systems.

### **3.5.4 System Accuracy and Precision**

In addition to EER% i.e. overall system performance – accuracy is also important to understand and specify in casework or analytical reporting. Ideally, results from ASRs should be accompanied with explanation and context regarding confidence or risk that the system could produce an incorrect result. For example, if a single speaker comparison is conducted a single LR value cannot represent

the inherent variability (of score output) that would naturally occur but which it is impossible to measure unless multiple comparisons are conducted (Morrison, 2010). Simply put an operator cannot know where the sole LR score they obtained is positioned with respect to overall variability if multiple comparisons were available. It is impossible to know whether a single score is higher than average, lower than average or an outlier. This can be better explained in the context of accuracy and precision if we imagine that a single comparison that we conduct falls as a point within a distribution curve (Figure 3.23).

**Figure 3.23:** Accuracy and Precision explanation  
From: Slideplayer.com/slide/7474261



### 3.5.5 Cost of Log Likelihood Ratio

The Cllr is a performance measurement that provides a metric of accuracy. It is particularly useful for evaluating systems with similar or low EER% and comparing how accurate they perform. Brummer and Leeuwen (2006) state that ‘a perfect recognizer (that makes no errors) will have zero loss, while all others have positive loss’ (2006: p.5). Cllr is also discussed in ENFSI standards as a useful measurement of accuracy: ‘The closer to value of Cllr is to zero, effectively the more accurate the system’ ENFSI (2015: p.26). Cllr is a measurement related to the applied probability of error (or APE) and to calculate it the APE is computed for a range of priors and considered with the LR output of the system. Both are plotted and the area of the difference calculated as the Cllr. The OWR Bio-Metrics system uses the Brummer and Leeuwen (2006) method for calculating the Cllr and this is used in the presentation of results for the experiments conducted in this thesis to discuss accuracy.

Recently, Cllr has gained popularity in validating system accuracy in conjunction with overall EER% performance and recommendations have been made for its incorporation into method validation (see Morrison, Thiruvaran and Epps, 2010; Hughes, 2014).

ENFSI also provide examples as to how Cllr could be incorporated into method validation/reporting and offers an example for an acceptable range. ENFSI provides an example as a possible validation criterion that the ‘Cllr for the method under evaluation should be smaller than 0.65.’ ENFSI (2015: p.29). Cllr threshold(s) should be applied to individual systems and in consideration of calibration – so it should be emphasised that this is just an example, rather than a direct recommendation. Broad recommendations (Hughes et al., 2019) are applied such that a Cllr of <1.00 is viewed as an acceptable level of accuracy – with values of Cllr >1 suggesting an ASR system may require adaptation of settings, or calibration, or that the audio itself is not of sufficient quality to obtain an accurate result. It is suggested that the application of Cllr, as a metric, is relatively recent and most useful – but the guidance for Cllr acceptability requires further clarification.

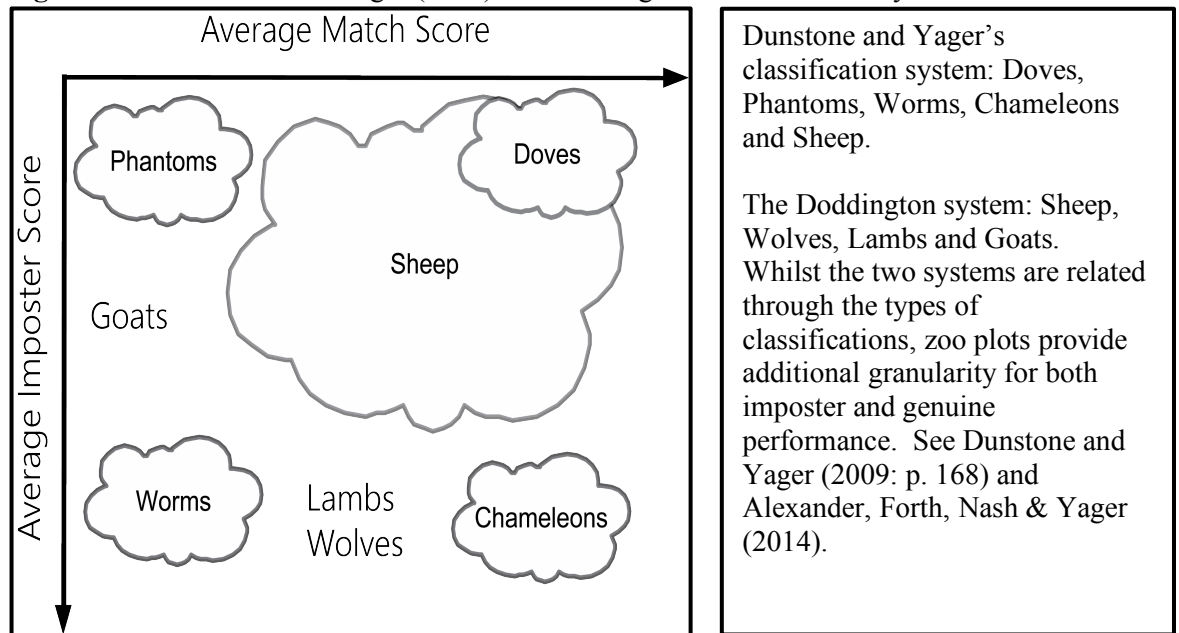
### **3.5.6 Zoo plots**

It would be convenient if all speaker models performed in a uniform way. Unfortunately, that is not the case. In analysing the baseline data (known results) on an ASR system Campbell (1997) described speakers as wolves or sheep dependent on tendency to false accept. Doddington et al. (1998) then applied the term ‘speaker menagerie’ and increased the classifying of speakers further to include sheep and goats (in addition to lambs and wolves). Doddington attributed an animal characteristic to each speaker as follows, loosely linking animal type to the way in which he felt they performed within a system.

- i. Wolves typically impersonate other speakers
- ii. Goats are difficult to identify
- iii. Lambs are easy to impersonate
- iv. Sheep describe the ‘normal’ distribution

Dunstone and Yager (2009) expanded on this idea and introduced new classifications. A visual summary of the two classification systems is presented below (Figure 3.24).

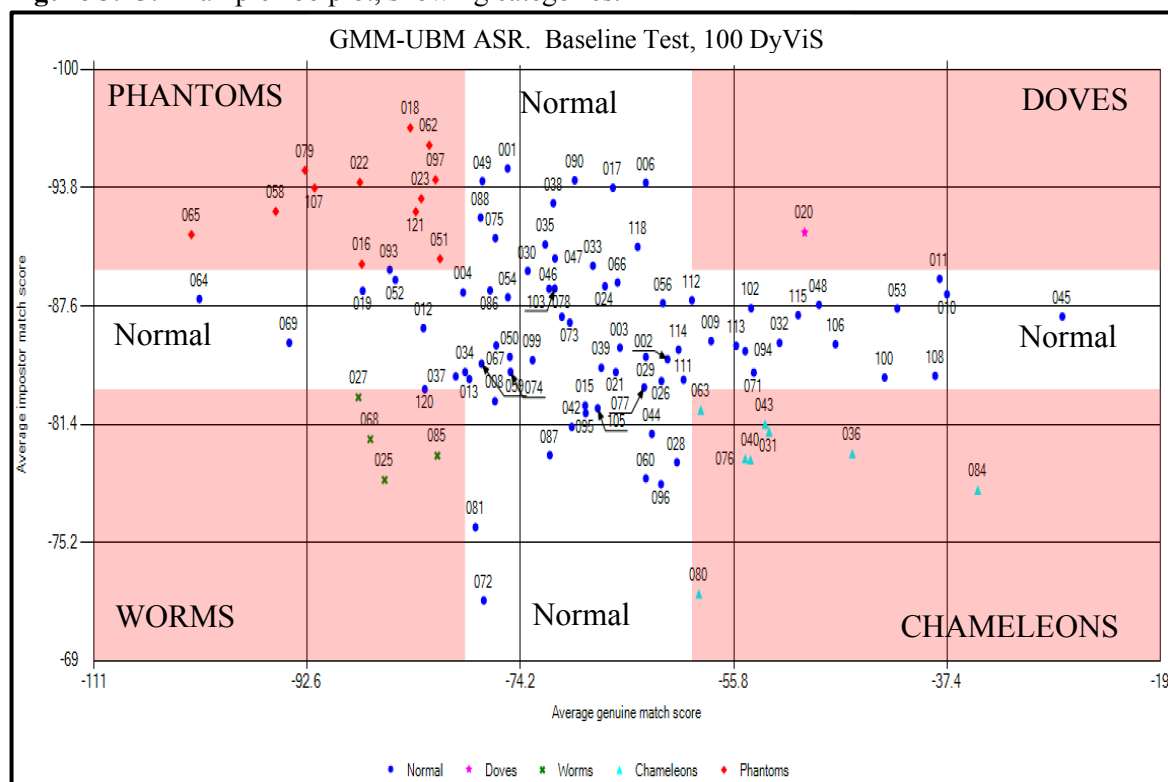
**Figure 3.24:** Dunstone and Yager (2009) and Doddington's classification systems



To better visualise the data and this system of classification Dunstone and Yager (2009) developed zoo plots (example in Figure 3.25). Zoo plots assign speakers to either normal or non-normal classifications on a two-tone x, y axis grid dependent on their performance. The x axis shows the mean likelihood ratio (LR) output from iVocalise or the log likelihood ratio (LLR) output from Vocalise for the same/matched speaker outcomes. The y axis displays the mean LR or LLR outcomes for the imposter/different speaker outcomes.

The dove, worm, chameleon and phantom categories are displayed in each of the four quadrants with normal distribution effectively forming the fifth classification in the white, central region. For the OWR Bio-Metrics zoo plot software, designed in consultation with Yager, classifications are calculated by taking the top and bottom 25% scores for both genuine and imposter matches to generate each of the quartiles.

**Figure 3.25:** Example zoo plot, showing categories.



Zoo plots are increasingly applied to examine candidate performance and stability of systems in many forensic fields, including face and fingerprint recognition (O’Conner et al., 2013). To summarise and place in the context of speaker comparison the zoo plot classifications are described by Dunstone and Yager (2009: p.161) as:

**Doves** are the best performers in a system. They produce high match scores against their speaker model and low match scores against the imposter models. To the ASR system dove speakers are easily recognisable and effectively stand out from the other comparisons completed.

**Chameleons** produce high match scores against their speaker model and high match scores against the imposter models. To the ASR system chameleon speakers appear similar to everyone.

**Phantoms** have low match scores against their speaker model and low match scores against the imposter models. To the ASR, system phantom speakers do not appear similar to anyone.

**Worms** are the worst performers in a system. They produce low match scores against their speaker model and high match scores against imposters. To the ASR system worm speakers are not easily recognisable and can be easily confused for other speakers.

**Normal** is the only classification to appear in both zoo plots and the Doddington system (sheep). This is the typical distribution.

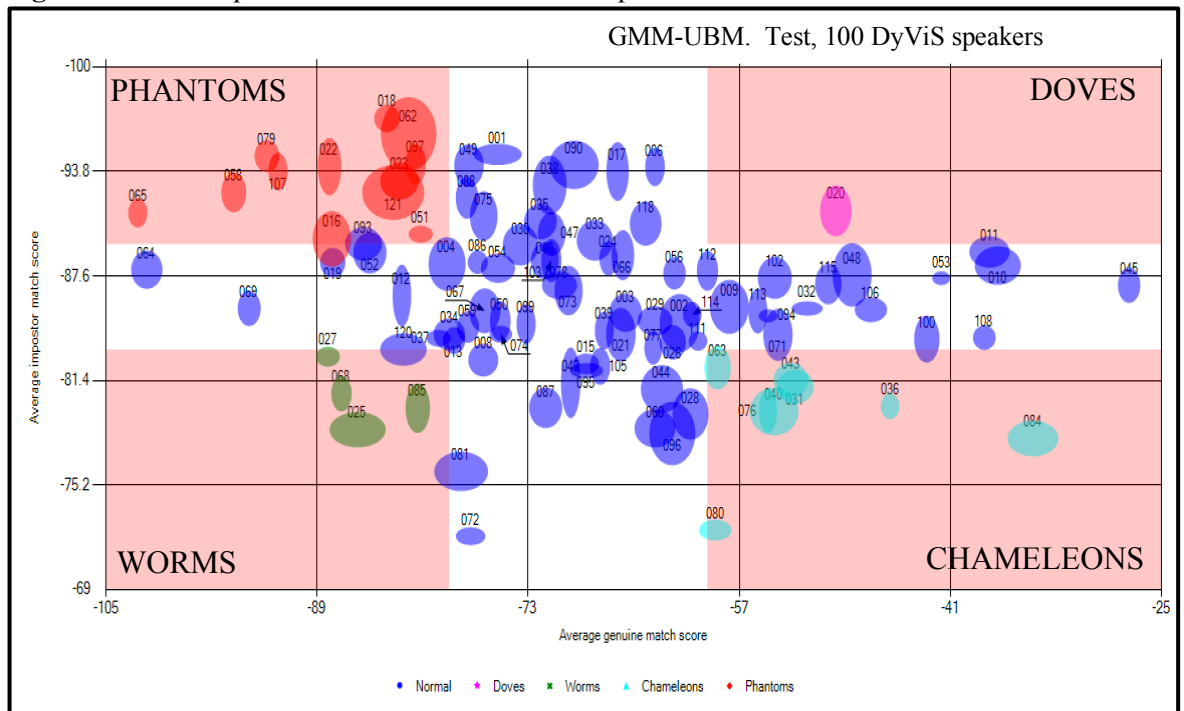
There are advantages and disadvantages to zoo plot analysis. They provide detail as to relative speaker performance in terms of how well a speaker can be distinguished against the others in the test and against themselves. Zoo plots can also be a useful tool for checking calibration and the relevance of normative data and skewed patterning was noticed during preliminary tests, particularly when DyViS was incorporated in the normative data (chapter 6 and Appendix D). Alexander, Forth, Nash and Yager (2014) recommend that: ‘Zoo plot analysis is done as speakers are added into a database, to help identify commonalities of speaker groups or algorithmic weaknesses of systems.’ (2014: p.1).

Another aspect of zoo plot analysis pertains to clustering although caution should be exercised in drawing definitive conclusions. Schnitzer et al. (2013: p.1) refers to clustering as ‘hubs’ - a natural cause of biometric comparison systems and that hubs contribute directly towards Doddington’s classifications (and subsequent zoo classifications). However, hubs are described as one symptom of near neighbour and average calculations for multiple similarity computations. Schnitzer’s study demonstrated that the more feature dimensions that are considered the greater the exaggerated effects of hubs and production of outliers (Schnitzer et al., 2013: p.5). So, whilst clusters, groups and hubs are important to examine, in themselves, zoo plot position cannot be fully conclusive in terms of causality and additional data and/or analysis is recommended to validate position and cause. As an example of this, in an early preliminary test, regionally accented speech data was added to DyViS accented speech (see 6.5.4 and Appendix D). In zoo plot analysis the accent speakers were observed to cluster in one quartile (phantoms). However, in that instance it could not be fully determined whether clustering was caused by the audio channel (different recording sessions) or the accented speech or both. The preliminary test and zoo plot results therefore influenced the experiments conducted, emphasizing the importance of constraining channel and intrinsic variability to avoid conflation. Note also that further research relating to zoo plot analysis is currently underway by Wang, Hughes and Foulkes (2019).

### **3.5.6.1 Zoo Plots and Inter/Intra Variability**

Inter-speaker variability describes the variation in speech between a speaker and other speakers. Intra-speaker variability describes the variation within a speaker’s speech. By default, single data points are displayed. However, Bio-Metrics can also display elliptical shapes to provide an indication of the degree of inter-speaker variability (distinction from other speakers) and intra-speaker variability (within speaker consistency) – see Figure 3.26. The additional zoo plot display option, for this variability, is referred to as ‘fat, thin, tall or short’ animals (Alexander et al., 2014).

**Figure 3.26:** Example of an OWR Bio-Metrics zoo plot with fat and thin animals



An ellipse shape (in either vertical or horizontal orientation) displays the intra or inter values for each speaker with the size of the ellipse/circle indicating the relationship to the mean. In summary, the mean of the standard deviation (for all speaker scores) becomes a circular unit of 1. Speakers with values larger than 1 are then referred to as fatter and/or taller in comparison to the other speakers in the test. Speakers represented by shapes smaller than the unit of 1 are referred to as thinner and/or shorter (Figure 3.26).

- i. Single unit circle = single unit = average (of this dataset)
- ii. Short and thin = low impostor variability scores, low genuine variability scores (Low inter, low intra variability)
- iii. Short and fat = low impostor variability scores, high genuine variability scores (How inter, high intra variability)
- iv. Tall and thin = high impostor variability scores, low genuine variability scores (High inter, low intra variability)
- v. Tall and fat = high impostor variability scores, high genuine variability scores (High inter, high intra variability)

The mean shape can be displayed in later versions of Bio-Metrics, as a circle, to provide a reference.

When expressing data using this visual representation an observation in preliminary tests was that intra and inter-speaker variability was not necessarily linked to classification. This is because the likelihood score and variability of that score are not linked variables (i.e. magnitude of LR and standard deviation from the mean) and it is therefore possible to have an animal of any width and height in any classification. In summary, zoo plots can provide indication of system performance

health, identify outliers and speakers that perform with similar scores (cohort groups/clusters) and also intra/inter speaker variability. Zoo plots enable a practitioner to visualise ASR speaker performance in a far more accessible and detailed way than single performance figures such as EER%. It is for those reasons that zoo plots were used extensively during the course of this thesis to examine ASR results.

### **3.5.6.2 Performance Measurements (False Accept Rate and False Reject Rate)**

The results tables in chapters 7-11 contain the following terms, which require explanation.

- i. H0 Mean: the average LR/LLR score for the hypothesis that two speakers compared are the same (genuine speaker match).
- ii. H1 Mean: the average LR/LLR score for the hypothesis that two speakers compared are not the same (imposter match).
- iii. H0 Standard Deviation (SD): this is effectively the measure of score spread for genuine match results. SD is the square root of the variance. Variance is calculated as the average of the squared differences from the mean.
- iv. H1 Standard Deviation (SD): the measure of score spread for imposter match results.
- v. FAR (False Accept Rate) and FRR (False Reject Rate). In determining system thresholds there is a trade-off between false accepts and false rejects. A well performing system obviously has very low false accepts and very low false rejects. In Bio-Metrics software (OWR), it is possible to represent this data by viewing the decision threshold on a sliding scale (from low to high) effectively decreasing the FAR at 0.01, 0.001 and 0.0001, which results in a corresponding increase in FRR. Viewing the relationship between FAR and FRR to this level of detail can be extremely useful, particularly when FAR is close to zero at 0.01.

## **3.6 Automatic Speaker Recognition Use Case Examples**

ASR systems are capable of completing hundreds of software comparisons per second. They do not fatigue and can produce a standardised set of results based on a defined set of algorithms and parameters/settings which are repeatable. When correctly operated and applied to high quality speech in sufficient quantity modern ASR systems can perform accurately (French et al., 2009). However, there is general agreement that accuracy can fall due to channel impairments (contaminants) including the effects of transmission and recording factors (French et al., 2009; Alexander, 2005). Nevertheless, ASR systems are growing in popularity for assisting with dealing with large volumes of speech data. This section outlines the differences between use cases as requirements vary significantly. Research conducted on the different application of ASR systems produced the following summary (Table 3.27).



**Table 3.27:** Typical examples of ASR use cases

<b>Example Sector</b>	<b>Question/use case</b>	<b>Type of ASR analysis</b>	<b>Comments</b>
<b>Healthcare, Banking, Insurance or Call Centre</b>	Is this Mr Smith on the telephone?  ASR assists with customer identity validation.	1 to 1 comparison. Relatively high security as used in combination with other data.	Prior expectation of customer identity, often text dependent. Compliant customer. Repeatable process if verification fails (including enrolment phase).
<b>Law Enforcement (example 1)</b>	Does this recording contain Mr Smith?	1 to 1	To investigative standard with some progression to evidence (auditory analysis underpinning).
<b>Law Enforcement (example 2)</b>	Does this recording contain Mr Smith or one of his associates?	1 to N	Non-evidential. To investigative standard, with some progression to evidence likely if other data assists with verification (auditory analysis by expert witness).
<b>Law Enforcement (example 3)</b>	Do these recordings contain any of our suspects?	N to N	For investigative purposes with some progression to evidence (auditory analysis underpinning).
<b>Forensic Practitioner</b>	As for Law Enforcement Agencies (LEAs)	1 to 1, 1 to N, N to N	Not yet approved for evidential purposes in the UK.

Differences in ASR application can determine the setting of threshold(s) based on the risk of incorrect outcome(s), the selection of normative data and settings pertaining to the mitigation of error. For example, a text dependent ASR system for telephone banking which completes a 1 to 1 speaker verification (i.e. questioned audio compared to speaker model from file) may have a very high threshold to limit false positives. Conversely an ASR applied in an investigative context on bulk data (perhaps on poorer quality audio and/or with lower net duration) may have a verification threshold set deliberately low to mitigate against false rejection. False positive and false reject outcomes have different repercussions such as incorrect inclusion or incorrect exclusion from an investigation. The trade off, of more false positive results requiring additional resources to validate against other data types, will be weighted against the risk of a criminal potentially evading detection.

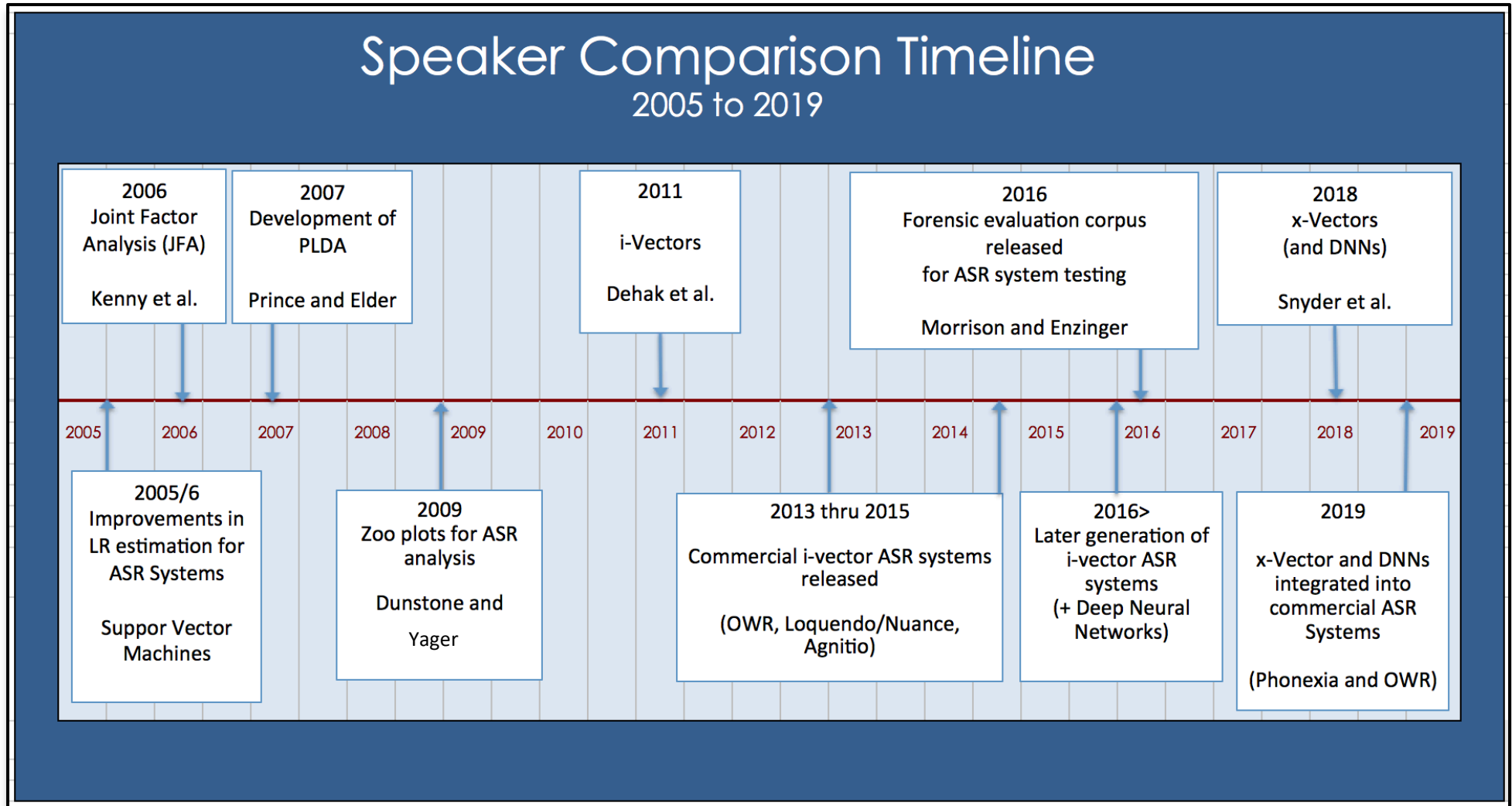
### **3.7 Summary**

In providing an explanation of technical terms and concepts this chapter highlights the complexity of the end-to-end process(es) from the speaker(s) through to the ASR/practitioner and the considerable variability that can be encountered throughout. For ASR systems to function to a high degree of accuracy throughout all these processes must occur successfully.

Obtaining full metrics from every section of the end-to-end signal chain is not possible and, to some extent, this influenced the scope of the experiments conducted in this thesis. It was determined that the five topics chosen - net duration, SNR, reverberation, frequency bandwidth and transcoding could be analysed effectively under controlled conditions and were likely to have the most significant extrinsic influence on ASR performance.

Finally, a new timeline from the research completed for chapter 3, is proposed below (Figure 3.28).

Figure 3.28: Speaker comparison timeline, evolution from 2005 to 2019



# Chapter 4 Research Questions

---

The initial research question was to establish the degree to which acoustic variability influences ASR performance under 5 conditions. The aim of this was to assist with informing ASR application in casework where a wide variety of acoustic conditions are commonly encountered. In improving the understanding of ASR performance on degraded audio, a main objective was to reduce errors and incorrect outcomes which could potentially have implications in terms of material presented to court, particularly in countries where ASR output is accepted as evidence.

A large number of experiments, including over 540 tests and creation of more than 16 million data points were completed, using a single research corpus (DyViS) recorded under highly controlled conditions to produce detailed metrics across the 5 conditions. Maintaining data consistency across analysis was considered important to avoid the conflation of variables that can occur when applying multiple corpora recorded under different conditions. This chapter provides a summary of the individual research questions addressed in each of the sets of experiments. Hypotheses are presented in each of the subsequent chapters.

## **Chapter 6: Preliminary Tests**

- Q1:** How should baseline be best established and what are the optimum ASR settings?
- Q2:** Are the zoo plot classifications of speakers similar for MFCC and LTFD?
- Q3:** Are zoo plot classifications affected by the technical features of the recordings, i.e. SNR and net speech duration, rather than just those features intrinsic to the voices themselves?

## **Chapter 7: Net Duration**

- Q1:** How does a state of the art i-vector/UBM, TV, LDA+PLDA system perform in comparison to a GMM-UBM system under low net duration speech conditions?
- Q2:** For the i-vector system, is performance degradation linear or are there any identifiable tipping points? If so, what are the optimum net duration settings for performance and net duration acceptability?
- Q3:** If 50 x speakers from the baseline test audio (i.e. 1m x 2 for 50 speakers) are compared with 50 speakers from the shorter duration test audio (i.e. 20s x2 per 50 speakers) is zoo plot position influenced by net duration when using 1m (baseline) speaker models for all 100 speakers?
- Q4:** In the very short duration results (e.g. 1-3s) is there any noticeable lexical/phoneme commonalities or spectrogram observations that explain zoo plot positioning for speakers who perform well (Doves)? Conversely, do the very poor performing speakers (Worms, Phantoms, Chameleons) exhibit high lexical divergence or any notable spectrogram observations?

## **Chapter 8: Signal to Noise Ratio (SNR).**

**Q1:** To what extent does decreasing the SNR influence ASR performance on modern systems and can any tipping points be identified?

**Q2:** Are speakers with lower existing SNR/poor vocal effort affected faster, in terms of performance degradation, as the SNR incrementally decreases? Conversely, are speakers with high SNR values more resilient to the addition of noise?

**Q3:** Does the addition of pink noise produce different results from the addition of white noise?

**Q4:** With regard to channel matching/mismatch, is there benefit from degrading the speaker models in line with the test audio or should the speaker models be held at the highest possible quality?

**Q5:** With regard to the degraded results, can processing plug-ins such as noise reduction and/or digital normalisation positively influence/restore ASR performance?

## **Chapter 9: Reverberation**

**Q1** How resilient are modern i-vector ASR systems to reverberation as opposed to the earlier GMM-UBM versions used in studies such as Castellano (1996) and Peer, Rafaely and Zigel (2008)? Further, how effective are session changes to an iVector ASR system, based on adapting the normative data (UBM, TV, LDA+PLDA), relative to one another?

**Q2** Under a given set of conditions, can we quantify the influence of reverberation on ASR performance? If so, are there any direct correlations with specific reverberation measurements such as RT60?

**Q3** Can the influence of reverberation be mitigated through:

- Matching conditions, i.e. RT60, for speaker model and test audio?;
- Adaptation or improvements to the normative data (i-vector/PLDA system) to potentially restore ASR performance?

## **Chapter 10: Frequency Bandwidth.**

**Q1** Does ASR performance noticeably improve relative to baseline when the frequency bandwidth is extended beyond telephony? If so, what is the optimum frequency bandwidth for ASR performance?

**Q2** Does an i-vector/UBM, TV, LDA + PLDA ASR system offer significant performance advantages over a GMM-UBM system when the frequency bandwidth is extended?

**Q3** Many ASR systems automatically down-sample audio files as they are imported, to a frequency bandwidth 0-4kHz (sample rate 8kHz). OWR Vocalise and iVocalise ASR software systems provide the operator with the opportunity to adjust the frequency bandwidth (minimum and maximum settings) for the MFCC feature extraction stage and allow the configuration of normative data. Can performance advantages therefore be found in terms of matching frequency bandwidth for speaker models and test audio?

- If we applied the same channel bandwidth limitation to both the questioned audio and speaker model, how would ASR performance vary against baseline?
- If iterative bandwidth degradation was applied to the test audio but wide band speaker models were used, how would ASR performance vary against baseline?

**Q4** If the frequency bandwidth is significantly reduced below that of standard telephony what implications would that have for ASR performance?

### **Chapter 11: Transcoding**

**Q1** How resilient are more modern i-vector/PLDA ASR systems to codec degradation in comparison with GMM-UBM systems?

**Q2** To what extent does ASR performance degrade when transcoding processes are applied to baseline data?

**Q3** How will compression codecs influence ASR performance?

**Q4** Can any operating thresholds be extrapolated relating to data compression rates which may assist with informing ASR use?

# Chapter 5      Equipment and Recordings

---

This chapter provides a brief overview of the equipment and research corpora which is consistent throughout all the experiments completed in this thesis. Additional detail is also provided within each chapter to document where materials and methods differ.

## 5.1 Software

### 5.1.1 Audio Applications and Scripts

Audio files were edited and analysed using the following software:

- i. Adobe Audition version 3.03 (2012) [Adobe.com/uk/products/audition](https://www.adobe.com/uk/products/audition)
- ii. Izotope RX Advanced, versions 3 (2012) through to 6 (2018) [Izotope.com](https://www.izotope.com)
- iii. Praat [Fon.hum.uva.nl/praat](https://fon.hum.uva.nl/praat).

These products were primarily selected due to the ease of operation and intuitive graphic user interfaces (GUI). In addition, all software was known to have undergone iterative updates over many years and were considered stable. Finally, in reference to the extremely high quality of batch export/transcoding required, testing completed by [Src.infinetwave.ca](https://www.src.infinetwave.ca) demonstrated that they were transparent in operation (did not further degrade or add artefacts) in comparison to other applications.

The above software was also validated to ensure that artefacts or additional variables/unwanted noise was not added. This was completed, for example, by analysing spectrograms to ensure the noise floor was not affected. In addition, null checks were undertaken - involving alignment of audio files in the time domain then inverting the phase of one and summing them together to check total phase cancellation (i.e. silence), see <https://www.soundonsound.com/techniques/phase-demystified>.

Care was also taken to ensure that audio file integrity was maintained throughout all the editing and processing stages. Dip sample checking (approximately 10-20%) was completed, applying auditory and spectrogram analysis (e.g. ensuring that additional noise and/or aliasing did not occur).

Using validated ground truth (or baseline) data has the benefit of knowing that, in each single comparison, exactly one of the speakers will match at least one of the test audio files. The audio files that formed the baseline data were edited to generate 30,000 reconcilable cross comparisons – by taking audio files from 100 speakers and effectively dividing them into 4 portions i.e. 1 speaker model (SM) and three test audio (TA) files per speaker. This then provided 29,700 imposter outcomes and 300 genuine speaker scores. Further details are provided in each of the chapters as the

test audio varies marginally for each experiment. The editing of SM and TA files was completed by hand. Automatic and semi-automatic diarisation software was tested but not regarded as suitable for the experiments due to the additional variability that they added.

To automatically split test files - e.g. for the net duration chapter - several Python ([Python.org](https://www.python.org/)) batch scripts were created to assist with generating multiple session data from the same speaker swiftly. The output, from the batch processes, were dip sampled (approximately 10-20%) to ensure accuracy and that the process itself did not contaminate the audio files. Awave software by FMJsoft ([Fmjsoft.com/awaveaudio](https://fmjsoft.com/awaveaudio/)) was used to complete the codec file conversions for the preliminary tests. The version used was 11.1.

## 5.1.2 Vocalise and iVocalise Software

The Vocalise and iVocalise ASR systems (Alexander et al., 2016) by Oxford Wave Research (OWR) [Oxfordwaveresearch.com/products/vocalise](https://oxfordwaveresearch.com/products/vocalise) were chosen for the research conducted in this thesis. They are similar in architecture and performance to other commercially available ASR systems.

The Vocalise, GMM-UBM, ASR software was available for the preliminary tests from 2012 onwards, in Beta version. The iVocalise, i-vector ASR system was available from 2015. More details, specifications and versions etc. are documented in Appendix G. These specific ASR systems were chosen for several reasons:

- The options and settings available to the user are extensive and enable a high level of system adjustment. This provided, for example, greater ability to analyse multiple types of ASR conditions and assist with determining whether the ASR can be adjusted to compensate for acoustic variability. It should be noted that many options, such as compiling complex normative sets and adjusting feature extraction settings, are not available or not as flexible on all commercial systems.
- Two generations of Vocalise ASR architecture (GMM-UBM and i-vector) were made available for assessment, providing a unique opportunity to test similar systems with different underlying methodologies.
- OWR offered unique insight into how their systems worked. This was evident through the provision of documentation, free and regular patch updates, responsive technical support and permission to baseline their ASR systems under difficult and complex conditions. Other commercial companies were approached but were unable (or unwilling) to provide this.



- In the Vocalise system (GMM-UBM) the normative set (UBM) can be relatively easily defined/compiled by the user in comparison to other commercial systems. Indeed, for some ASR systems changing the normative data is strongly discouraged in preference to a default set which is often of unknown compilation.
- Various system options are available to the user which are not available on other ASR systems. For example, in Vocalise, options were also available to the user for exploring different methods of feature extraction, such as long-term formant distribution (LTFD). This unique feature extraction method was explored in the experiments and these are further explained in the relevant sections.

The iVocalise ASR system uses i-vectors for statistical modelling (see 3.4.5.2). For an i-vector system to work successfully the normative dataset requirement is much larger than for GMM-UBM. OWR provided assistance in compiling normative data for the i-vector system because a much more complex set of UBM, TV, LDA and PLDA enrolment is completed – using one, very large set of normative .wav files. This process is referred to by OWR as a ‘session’. Similar to Vocalise, the iVocalise system also allows for user configuration and parameter changes which are often unavailable to users of other ASR systems and these are documented in the relevant sections.

Both iVocalise systems output a comma separated value (.csv) file which contains all the output data from the comparisons completed (e.g. successful comparisons completed and LR or LLR score output). The iVocalise systems are commercially available and widely considered to be comparable to other state-of-the-art ASR systems in terms of performance. This was recently tested in a set of studies which examined different ASR systems and further information can be found at Morrison and Enzinger (2019) and Kelly et al. (2019).

### **5.1.3 Bio-Metrics Software**

The output .csv files from Vocalise and iVocalise were examined using OWR software Bio-Metrics 2011a [Oxfordwaveresearch.com/products/Bio-Metrics](http://Oxfordwaveresearch.com/products/Bio-Metrics). Bio-Metrics software exploits the iVocalise output files to provide metrics such as Cllr, EER% and can complete a wide variety of charting and graphing functions such as LR plot, and zoo plot to assist with system performance analysis. Recent versions of Bio-Metrics (late 2018 onwards) can also complete score system fusion. It should be noted that this function was not available at the time that the experiments were completed.

## **5.2 Summary of Hardware**

Two, standard build, Apple computers (A1286 and A3198) were used throughout this thesis. These were used for all audio editing, analysis and for running the Parallels VMWare software

[Parallels.com/uk/landingpage](https://parallels.com/uk/landingpage). Parallels is a virtual PC that provides access to Windows OS and the OWR suite. A standard Dell XPS15 laptop was also used, primarily to validate that the Parallels software was transparent in (audio) operation. Audio files processed using the Parallels VMWare system and the Dell XPS15 laptop were compared and determined to be technically identical (i.e. Parallels did not add artefacts or degrade the audio files). Complete audio file integrity, in Parallels VMWare, was also confirmed through direct correspondence (see Appendix L).

Beyer DT990 Pro headphones were used for listening, [Europe.beyerdynamic.com/dt-990-pro.html](https://Europe.beyerdynamic.com/dt-990-pro.html). The frequency response (5Hz to 35kHz) makes them particularly suitable for monitoring and detailed audio analysis. The Avid/Digidesign Mbox 3 series audio interface (USB) was used in preference to internal PC soundcards, which were often found to introduce small amounts of mains hum or noise into the headphone socket output. The Mbox series is now discontinued but details can be found at: [Akmedia.digidesign.com/support/docs/Mbox\\_Technology\\_Guide\\_70405.pdf](https://akmedia.digidesign.com/support/docs/Mbox_Technology_Guide_70405.pdf).

### **5.3 Speech Corpora**

The Dynamic Variability in Speech corpus or DyViS (Nolan and McDougall et al., 2009) features 100 male speakers between the ages of 18 and 25. All speakers are classified as Southern, Standard, British, English (SSBE). A number of speaking tasks were undertaken by participants and the free speech, a simulated police interview (task 1) and a simulated telephone conversation (task 2), were selected as the most forensically realistic.

The task 1 (microphone, 44.1kHz sample rate, 16bit depth) and task 2 (telephone, 8kHz sample rate, 16bit depth) data were selected for this thesis to reflect typical casework conditions. The DyViS corpus was also selected due to the overall high quality of the recordings and the strictly controlled conditions in which they were created in addition to the metadata available to assist analysis. It was determined that any inherent variability or small amounts of channel variation within the corpus would become a part of baseline ASR performance i.e. acoustic degradation (contaminants and inhibitors) to be applied to the baseline recordings. This underlying methodology was common to all experiments conducted.

Additional corpora were used to provide bespoke normative data where the default normative sets were unsuitable. Further details are provided in the relevant chapters.

# Chapter 6 Preliminary Testing

---

This chapter summaries the preliminary tests conducted prior to the main research experiments. The chapter is provided as a record to demonstrate how the methodology and scope of the subsequent research experiments was established and how the baseline data or ground truth, common to all the subsequent experiments conducted, was obtained. At the time the preliminary tests were completed the iVocalise (i-vector) ASR was not yet available.

## 6.1 Objectives

To provide accurate output from the experiments it was determined that the baseline performance of systems (EER%) should be reflective of a high performing state of the art and fully optimised ASR system. The preliminary tests therefore ensured that ASRs were correctly set, that the corpora and editing points were suitable and the selection of normative data was effective. Objectives were defined as:

- i. Familiarisation with Vocalise ASR system operation;
- ii. Testing the ASR feature extraction methodologies (i.e. MFCC, LTFD) and assessing performance differential (if any);
- iii. Preparation of speaker models (SM) and test audio (TA) files for baseline data;
- iv. Selection and preparation of normative data;
- v. Establishing if any technical (acoustic) or intrinsic variability could be determined within the DyViS corpora which could influence results from further acoustic variability tests;
- vi. Gaining familiarity with Bio-Metrics software to measure performance i.e. zoo plots, LR plots, EER% graphing;
- vii. Completing baseline experiments under controlled conditions and adjusting ASR settings to inform ASR performance reflective of a state-of-the-art system;
- viii. Providing assurance in terms of validating methodology, defining research experiments and determining scope.

## 6.2 Questions

The following research questions (Q) were set with associated hypotheses (H).

**Q1: How should baseline be best established and what are the optimum ASR settings?**

**H1:** In applying current research methodology a corpus recorded under carefully controlled condition should be used. SM and TA files should be carefully edited. Known performance outcomes should be attained – i.e. true positive (TP), true negative (TN), false positive (FP) and false

negative (FN). ASR settings should be adjusted to optimise performance and establish EER% performance figures for MFCC and LTFD extraction methods (referred to as ‘modes’ or ‘engines’). Testing will establish this.

**Q2: Are the zoo plot classifications of speakers similar for MFCC and LTFD?**

**H2:** It is hypothesised that the zoo plots are likely to show some performance variation between different engines as they are based on different measurements and therefore statistical speaker models. However, it is not known to what extent they will vary and the difference in EER% between the two systems will be an important element of the preliminary tests to establish which will be more effective to use in the main research experiments.

**Q3: Are zoo plot classifications affected by the technical features of the recordings, i.e. SNR and net speech duration, rather than just those features intrinsic to the voices themselves?**

**H3:** The corpus was recorded under highly controlled and consistent conditions e.g. microphone gain and position, sample rate, bit depth and room. It is therefore suggested that poorer performing speakers (i.e. high ASR imposter match scores and/or low genuine match scores) may not necessarily equate directly to technical features – since those are relatively uniform across the corpus. Nevertheless, outlying speakers which are classified in more extreme zoo plot positions could exhibit certain technical features such as those which are likely to vary across the corpus (e.g. SNR linked to vocal effort). Examination of intrinsic factors, such as voice quality or the addition of accented data could assist with explaining zoo position causality too and so experimental tests should also be conducted using additional (VQ/VPA) data.

## **6.3 Data Preparation and Materials**

The Dynamic Variability in Speech (DyViS) corpus (Nolan and McDougall et al., 2009) was selected for use in the experiments and permission was granted for use. DyViS features 100 male speakers between the ages of 18 and 25 recorded under controlled conditions (spontaneous speech). All speakers are classified as Southern, Standard, British, English (SSBE). The task II (telephone channel) speech files were edited and the following audio data was removed:

- i. The interlocutor/interviewer
- ii. Overlapping speech, i.e. the speaking and interlocutor speaking simultaneously
- iii. Any dial tones, beeps, GSM interference, clicks, crackles, distortion or clipping
- iv. Signal drop outs, silent pauses
- v. Non-speech sounds (coughs, breathing, sighs etc)
- vi. Any rustling, movement or environmental noise

The maximum net quantity of speech was obtained. This was edited into a speaker model (SM) and multiple test audio (TA) files per speaker from the same session to limit channel variability and prevent the conflation of variables. ASR tests were conducted examining different SM lengths. Whilst large differences in performance were noted at the <1m SM net duration point, only a very negligible differential in EER% performance was noted between the 1m and 3m duration lengths.

MFCC GMM-UBM. EER 1.24%: 1m SM

MFCC GMM-UBM. EER 1.01%: 3m SM

It was assessed that ASR performance was acceptable at 1m (SM) net duration which then provided enough material to provide multiple files for the TA for all speakers. This test informed the scope for the chapter on net duration, to further examine sub 1m SM and TA performance.

To provide sufficient speech material to inform both the SM and TA material it was therefore determined that edit points should be made in the following manner:

- i. First minute of net speech = SM
- ii. Second minute of net speech = TA 1
- iii. Third minute of net speech = TA 2
- iv. Remaining material (variable length files containing residual) = TA 3

This process was applied to each of the 100 Speakers – i.e. SM (100) and TA files (3 x 100) were created. This then provided 30,000 cross comparisons i.e. 29,700 different speaker/true negatives and 300 same speaker/true positives. It was noted that intra-speaker variability could be better measured with multiple session audio (of the same channel conditions). However, this option was unavailable within the DyViS corpus, unless introducing additional variability pertaining to sample rate, bit depth and codec through the addition of DyViS task 1 data (mock interview). To maintain channel consistency this was therefore not completed.

The OWR Vocalise system (GMM-UBM with options for MFCC and LTFD feature extraction) used was build 1.5.0.1190. Symmetrical testing is an option in Vocalise and this was selected to further improve performance – this effectively reverses the status of SM and TA to establish mean score values and is useful when net duration differs (i.e. in most cases). The OWR Bio-Metrics software used was build 1.4.0.597.

### **6.3.1 User Mode**

Note also that Vocalise 1 included the option to import hand annotated data (e.g. formant data). This is referred to as the ‘user mode’. The user mode was not used in the experiments, to limit variables.

## 6.4 Selection of Normative Data

The OWR Vocalise ASR provides the user the option of configuring normative data. It is widely understood that normative data should not contain the same speakers as the question/test audio files and speaker model(s) due to result distortion. It was determined that a bespoke normative data set was required. A normative dataset was constructed specifically to reflect the demographic of the speakers in the trials i.e. SSBE and male aged 18-25. The Speech Obtained in Key Environments (SPOKE, 2015) corpus was used for both MFCC and LTFD GMM-UBM experiments. SPOKE contains approximately 200 speakers (UK English) recorded using 8 different microphone types. The telephone (i.e. GSM transcoded/far channel) data was selected to best reflect task 1 in DyViS (i.e. high channel similarity). To ensure high normative relevance to DyViS the speakers for the normative data were selected from a similar speaker demographic to the test material (SSBE, male and 18-25) and of similar net speech duration as the SM and TA. The process for enrolling the normative data for Vocalise GMM-UBM, as defined by OWR, was followed.

It was noted that to achieve an EER of 1.2% (MFCC) and 6.02% EER (LTFD) in Vocalise a normative set was applied which contained very low numbers of speakers (less than 100) and additions appeared to make no further improvements to EER%. It was somewhat surprising that an EER% could be so low (and performance so high) using such a very small normative set. Although not conclusive in itself - this supported research on GMM-UBM normative data by Hasan and Hansen (2011) as noted in 3.4.7.

Different normative sets were tested, for example using material from SPOKE which did not reflect the SM and TA, other accented data and even DyViS. The results were often very poorly skewed (zoo plot) for both MFCC and LTFD engines and EER% raised significantly. Output was deemed useless - confirming the importance of separation of data between UBM and test/questioned audio (see Appendix D). In summary, this test demonstrated that there is a clearly a strong relationship between the normative data and individual speaker performance in addition to the overall system performance (EER%).

### 6.4.1 Additional Data

Stevens and French (2012) examined the voice quality of the 100 DyViS speakers in detail. Note that the voice quality settings and scores established for DyViS speakers by Stevens and French (2012) was recently superseded by a definitive set represented in San Segundo et al. (2018). However, at the time of conducting these preliminary tests, the Stevens and French estimations were all that was available. Stevens and French (2012) adapted previous methods to assess voice quality using an adjusted version of the Vocal Profile Analysis (VPA) scheme developed by John Laver (Laver, 1968; 1975; 1979; 1980; 1991). Each of the speakers was scored using a subjective six-

point scale (marked 0 to 5) for 34 vocal settings and a VQ data grid was produced. This was used to provide score indicators for each speaker. Additional grids were created for this thesis to analyse the distance from the mean for each VQ score (see Appendix B). VQ data was then examined in relation to the MFCC and LTFD zoo plot classifications.

## 6.5 Preliminary Test Results

The results from the preliminary tests are presented with associated observations.

### 6.5.1 Automatic Speaker Recognition Settings and Equal Error Rate Results

As expected, both MFCC and LTFD engines performed relatively well on the baseline data. Optimum EER points were established on the GMM-UBM systems as 1.244% (MFCC) and 6.022% (LTFD) shown in bold in Table 6.1.

**Table 6.1:** Summary of preliminary EER% results. Vocalise, DyViS, telephone channel

<b>Feature Extraction 'Engine'</b>	<b>UBM</b>	<b>Extraction settings</b> * Number of filters (see 3.4.3) **Default number of Gaussians is 32	<b>EER %</b>
LTFD	Type A SSBE UBM	F1, F2, F3 Default Gaussians**	8.686
LTFD	Type A SSBE UBM	F1, F2, F3, Default Gaussians**	7.483
LTFD	Type A SSBE UBM	F1, F2, F3, F4 12 Gaussians	7.737
LTFD	Type A SSBE UBM	F1, F2, F3, F4 24 Gaussians	6.308
<b><u>LTFD</u></b>	<b><u>Type A SSBE UBM</u></b>	<b><u>F1, F2, F3, F4 Default Gaussians**</u></b>	<b><u>6.022</u></b> <b><u>(optimum)</u></b>
MFCC	Type A SSBE UBM	Default Gaussians** Default with Cepstral Mean Subtraction (CMS)	4.991
<b><u>MFCC</u></b>	<b><u>Type A SSBE UBM</u></b>	<b><u>Default Gaussians** 13 filters*</u></b>	<b><u>1.244</u></b> <b><u>(optimum)</u></b>
MFCC	Type A SSBE UBM	Default Gaussians** 12 filters*	1.8468
MFCC	Type A SSBE UBM	Default Gaussians** Default plus Delta Delta	7.225
LTFD	DyViS [100]	Results Null: Normative data pollution	N/A
MFCC	DyViS [100]	Results Null: Normative data pollution	N/A

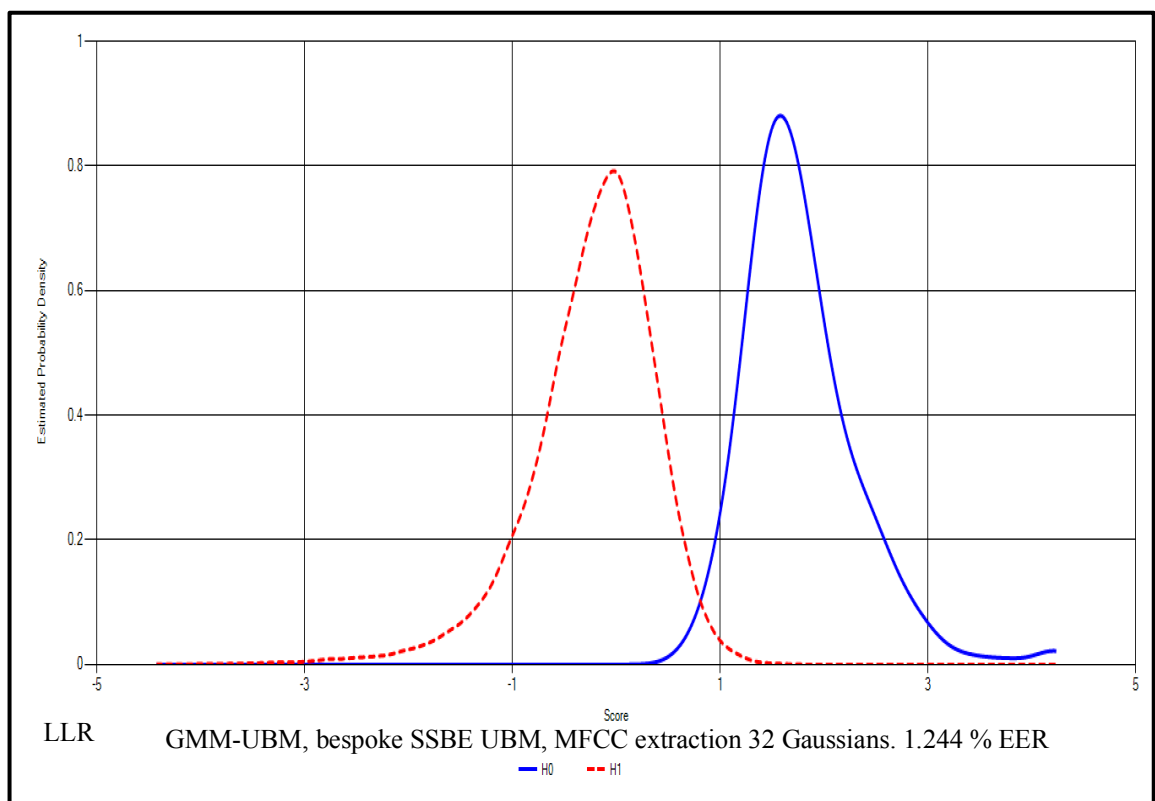
EER% performance varied and a link between number of gaussians and formants extracted was noted. A difference between LTFD and MFCC EER% performance was established and as predicted. Both these observations were also independently confirmed in Jessen, Alexander and Forth (2014).

Differences between LTFD and MFCC are likely due to the additional data captured by the MFCC process in comparison to formant values alone (LTFD engine). This was referenced in Rose (2013: p.84) who stated ‘there is potentially more information in a cepstral than a formant comparison’.

## 6.5.2 Cepstral and Formant System Comparison

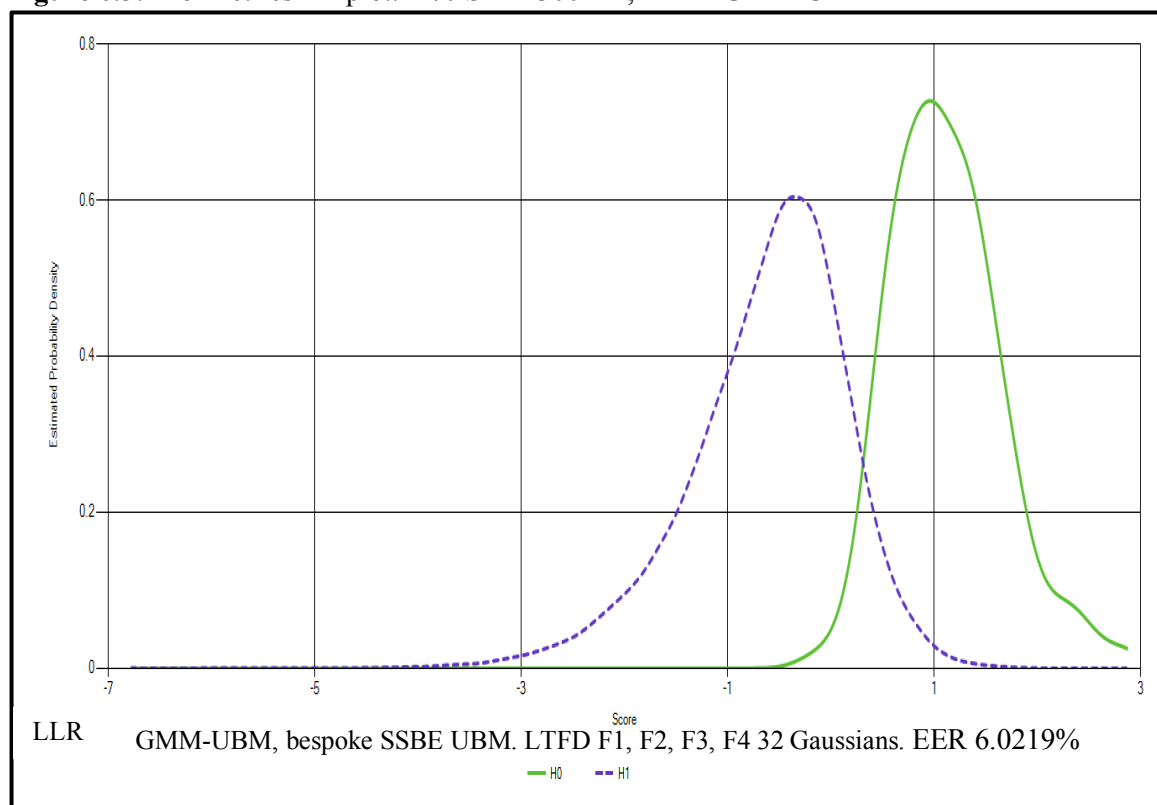
The OWR Vocalise ASR outputs a .csv file for analysis in OWR Bio-Metrics software. Two example LR Plots are presented below showing baseline tests using the optimum settings for the MFCC and LTFD engines (Figure 6.2 and Figure 6.3).

**Figure 6.2:** Bio-Metrics LR plot. 100 SM x 300 TA, MFCC GMM-UBM





**Figure 6.3:** Bio-Metrics LR plot. 100 SM x 300 TA, LTFD GMM-UBM



Lower scores were observed for the LTFD engine overall and score distribution separation (between same speaker and different speaker comparisons) was noted to be poorer in comparison to the MFCC engine.

### 6.5.3 Zoo Plot Analysis

Using zoo plot analysis it was observed that the MFCC engine produced speakers with marginally more phantom and worm classifications. This prompted the requirement, in further experiments, for additional metrics such as cost of likelihood ratio (Cllr) which was integrated into later versions of Bio-Metrics. As expected, Doves were greater in number for the MFCC system. Since both the MFCC and LTFD engines employ broadly similar extraction methods on the same data some commonalities in classifications were expected in terms of zoo placement. This was demonstrated in the results (Tables 6.4 and 6.5) where 8% of speakers appeared in the same zoo plot quadrant for both MFCC and LTFD extraction engines.

**Table 6.4:** Vocalise ASR, MFCC. EER 1.2441%: Zoo plot categories by speaker number

<b>Doves</b>	<b>012</b>	<b>047</b>	<b>008</b>	<b>071</b>	<b>038</b>	<b>049</b>	<b>020</b>	<b>003</b>		
<b>Chameleons</b>	<b>087</b>	<b>100</b>	<b>044</b>	<b>074</b>	<b>030</b>	<b>090</b>	<b>076</b>			
<b>Worms</b>	<b>025</b>	<b>063</b>	<b>107</b>							
<b>Phantoms</b>	<b>018</b>	<b>058</b>	<b>093</b>	<b>077</b>	<b>037</b>	<b>103</b>	<b>033</b>	<b>024</b>	<b>080</b>	<b>040</b>

Speakers classified identically in both MFCC and LTFD tests are highlighted.

**Table 6.5:** Vocalise ASR, LTFD. EER 6.0219%: Zoo plot categories by speaker number

<b>Doves</b>	<b>051</b>	<b>033</b>	<b>086</b>	<b>003</b>	<b>111</b>				
<b>Chameleons</b>	<b>066</b>	<b>076</b>	<b>030</b>	<b>015</b>	<b>100</b>	<b>087</b>	<b>050</b>		
<b>Worms</b>	<b>035</b>	<b>059</b>							
<b>Phantoms</b>	<b>018</b>	<b>093</b>	<b>042</b>	<b>024</b>	<b>054</b>	<b>053</b>			

The lower overall performance of the LTFD engine in the preliminary tests, the lack of possibility for including it in most ASR systems and the introduction of i-vector ASR systems at the time - was taken as grounds for not using it in the main experiments.

### 6.5.4 Voice Quality and Accent Data

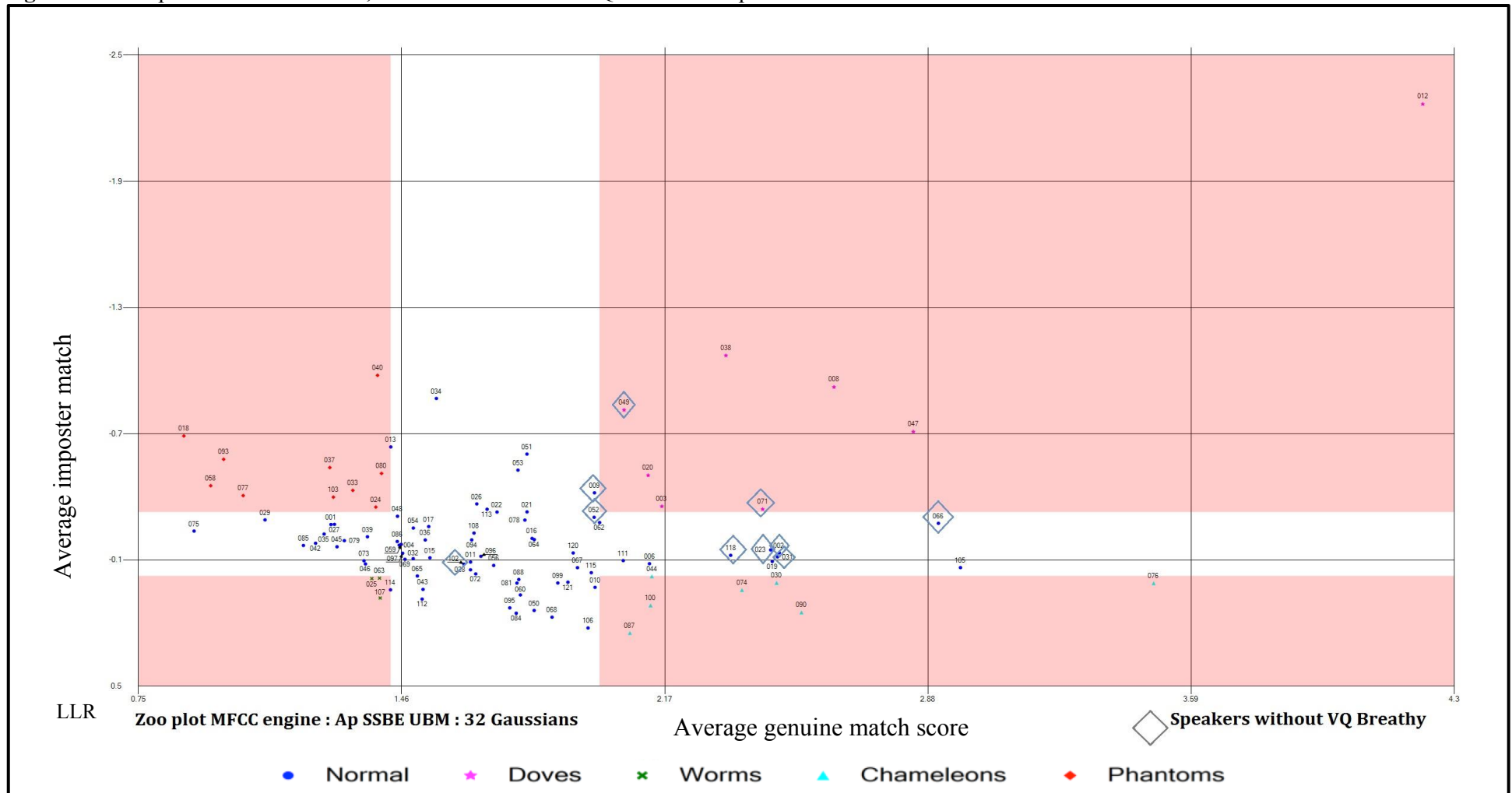
Voice quality data was provided, for DyViS speakers, from research conducted by Stevens and French (2013). A later research paper by San Segundo et al. (2018) re-examined voice quality for DyViS speakers. This was conducted in the context of inter-rater consensus for VPA where it was found that this was achievable within the group of three experts completing the method as outlined in their research.

Analysis of the Stevens and French (2012) data was completed and, using their scores for VPA, new tables were created which re-scored speakers as to standard deviation (see Appendix B) – i.e. distant from mean for all 100 speakers and ‘rarity’ of a given VQ feature (within the set of 100 speakers). This highlighted speakers which had an above average score for any given voice quality criteria. Zoo plot position was then examined in reference to VQ. Some clustering in regards to zoo plot position appeared evident for a small number of VQ features (e.g. Figure 6.6 and Figure 6.7). Additional examples of the zoo plots generated in relation to VQ data analysis are also provided in Appendix H.

In summary the subjective nature of the underlying VQ data suggested that, whilst some potential correlations with zoo plot position were observed, further research was required to establish which criteria were contributing to position and it was determined that this was outside the scope of the subsequent experiments.

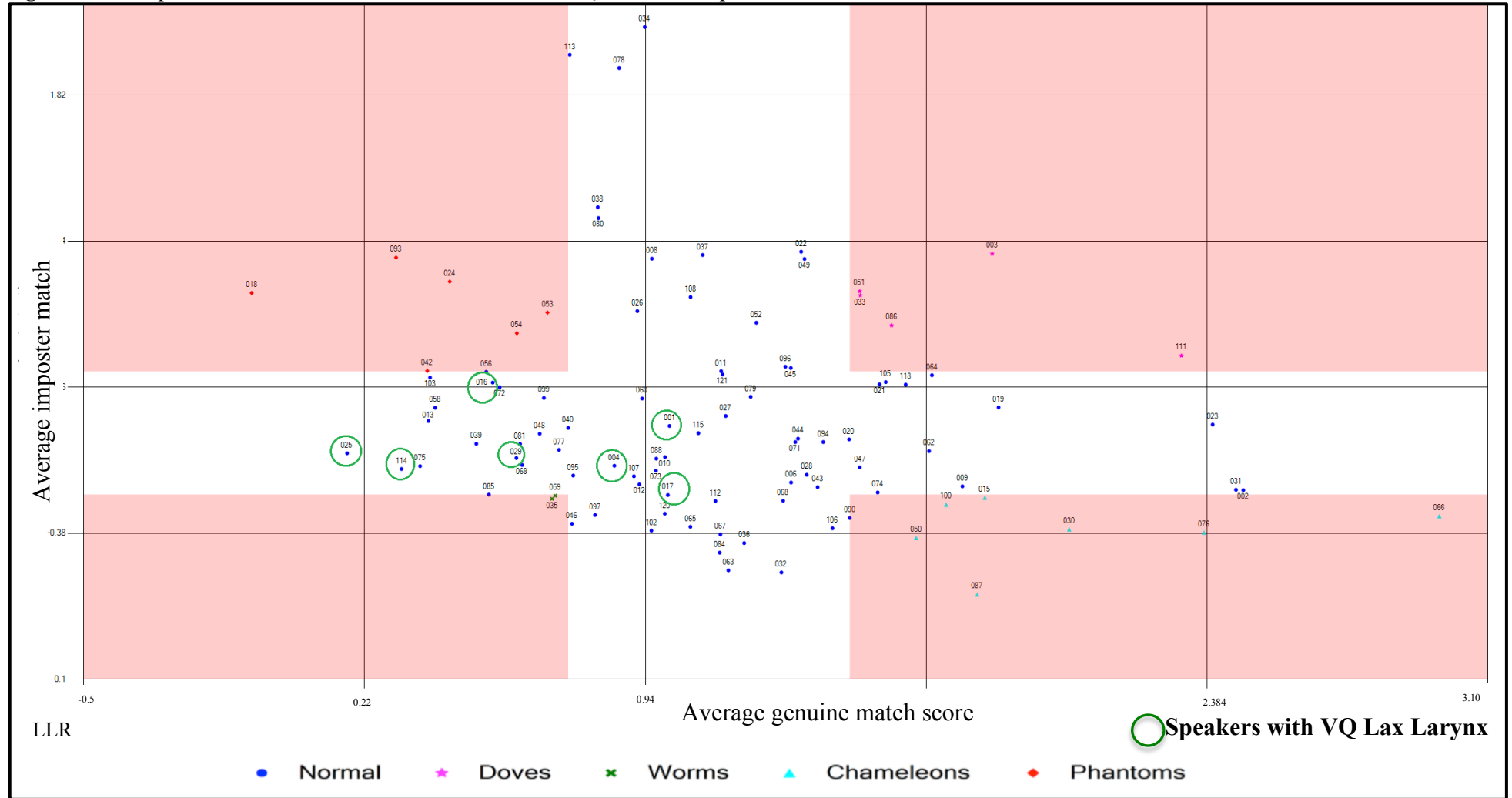
Finally, a brief test was completed examining zoo plot position and the addition of Pakistani and Yorkshire accented speakers (i.e. SM and TA) with the same telephone characteristics as the SSBE accented DyViS data. Clustering of accented data was observed in the zoo plot positioning. However, this only applied when normative data was selected using DyViS (i.e. skewed results negated the significance of position) see Appendix D.

Figure 6.6: Zoo plot 100 SM x 300 TA, GMM-UBM MFCC. VQ Data 1 Example



Clustering of speakers without VQ Breathy noted (right side of zoo plot).

Figure 6.7: Zoo plot 100 SM x 300 TA, GMM-UBM LTFD. VQ Data 2 Example

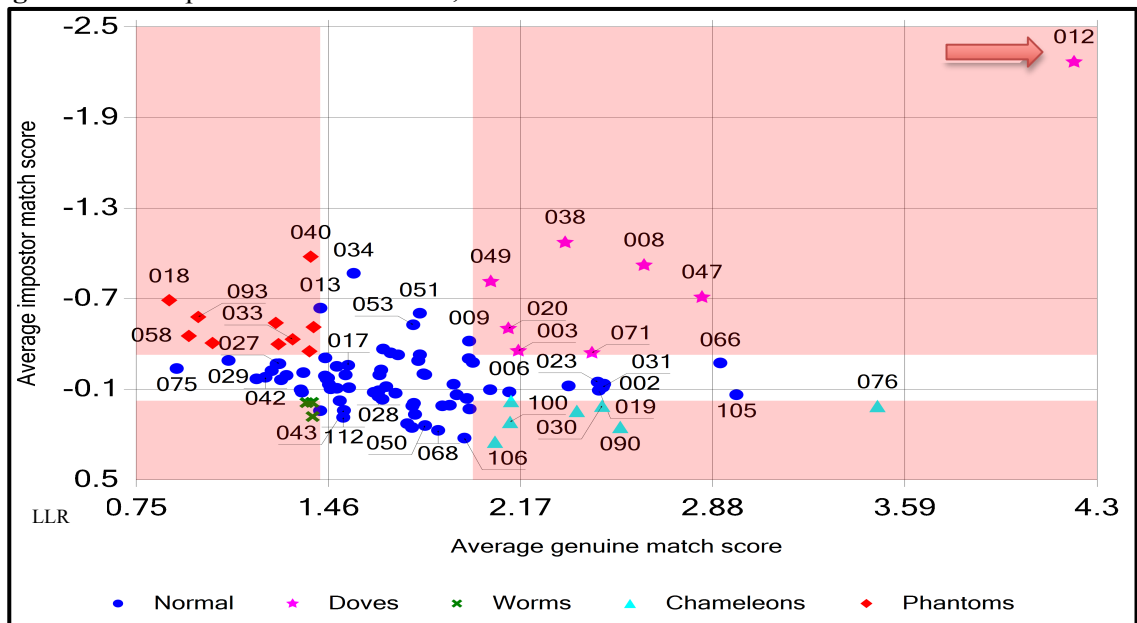


Clustering of speakers with VQ Lax Larynx noted (left side of zoo plot).

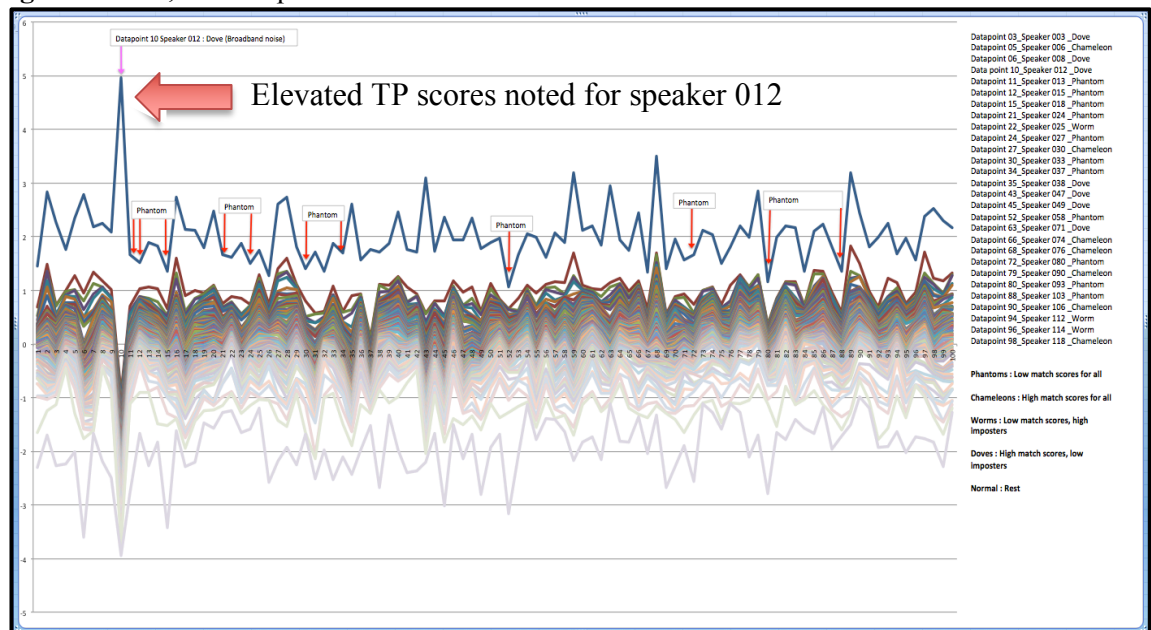
## 6.5.5 Additional Analysis (Speaker 012)

Since SM and TA files were extracted from the same corpus it was expected that speaker performance would be relatively consistent and zoo plot examination would yield little in terms of positioning with respect to technical quality. However, that was not the case. Zoo plot and spreadsheet analysis (Figure 6.8 and Figure 6.9) showed speaker 012 as an outlier dove by a wide margin (MFCC engine). Speaker 012 scored very high genuine match LR scores and very low imposter match LR scores. This position suggested that the speaker had either an extremely distinctive voice (to the ASR) or another variable was influencing speaker performance and zoo plot position. Unusually, speaker 012 did not appear in the same quadrant for the LTFD engine as the MFCC engine (Vocalise GMM-UBM).

**Figure 6.8:** Zoo plot 100 SM x 300 TA, GMM-UBM MFCC



**Figure 6.9:** 30,000 comparisons MFCC Vocalise. Blue line shows TP scores



Further analysis was completed using a spectrogram to view frequency content (see Appendix K). It was noted that the recording used in both the speaker model and the test sample, from the same session, contained 50Hz mains hum with associated harmonics (horizontal lines, fixed frequency). On re-examining both speaker 012 files and the remaining 99 speakers, this noise was not present for any other speaker. In understanding ASR feature extraction (MFCC engine) a plausible explanation for why this speaker produces very high match scores (and very low imposter scores) was therefore probably caused by noise not present in either the speaker model or test audio for any other speaker. I.e., speaker 012 is effectively easy for the ASR to distinguish due to non-speech (constant) values extracted. On adapting the zoo plot view to examine intra and inter-speaker variability (in relation to the mean) speaker 012 displayed as a ‘tall and thin’ speaker point (3.2.6) (Alexander et al., 2014) in comparison to other speakers (i.e. very high inter-variability, very low intra-variability). A plausible explanation is that noise is present within the 3 test audio files and speaker model not found in any other file.

Whilst the EER% was elevated, and performance therefore lower, the LTFD engine appeared to provide results more robust to the mains hum noise (speaker 012 not elevated). This is likely due to formant values estimated from mean values, which are effectively tracked throughout the audio file, rather than a full MFCC feature extraction (i.e. speech + noise). Speaker 012 was therefore classified as normal on the Zoo plot pertaining to the LTFD results. These speaker performance characteristics demonstrate the risk of acoustic variability – specifically the influence of noise when using MFCC extraction. It cannot always be assumed that ASR performance is based solely on the speech within the file. In summary, this preliminary test analysis highlighted the importance of examining the technical quality of audio, applying zoo plots to inspect ASR and speaker performance and the utility of spectrograms to examine acoustic variability. For completeness, speaker 012 was not deleted from the corpus. However, the DyViS Type I (interview) data was preferred for subsequent experiments due to the extended frequency bandwidth and absence of mains hum.

### **6.5.6 Signal to Noise Ratio Test Results**

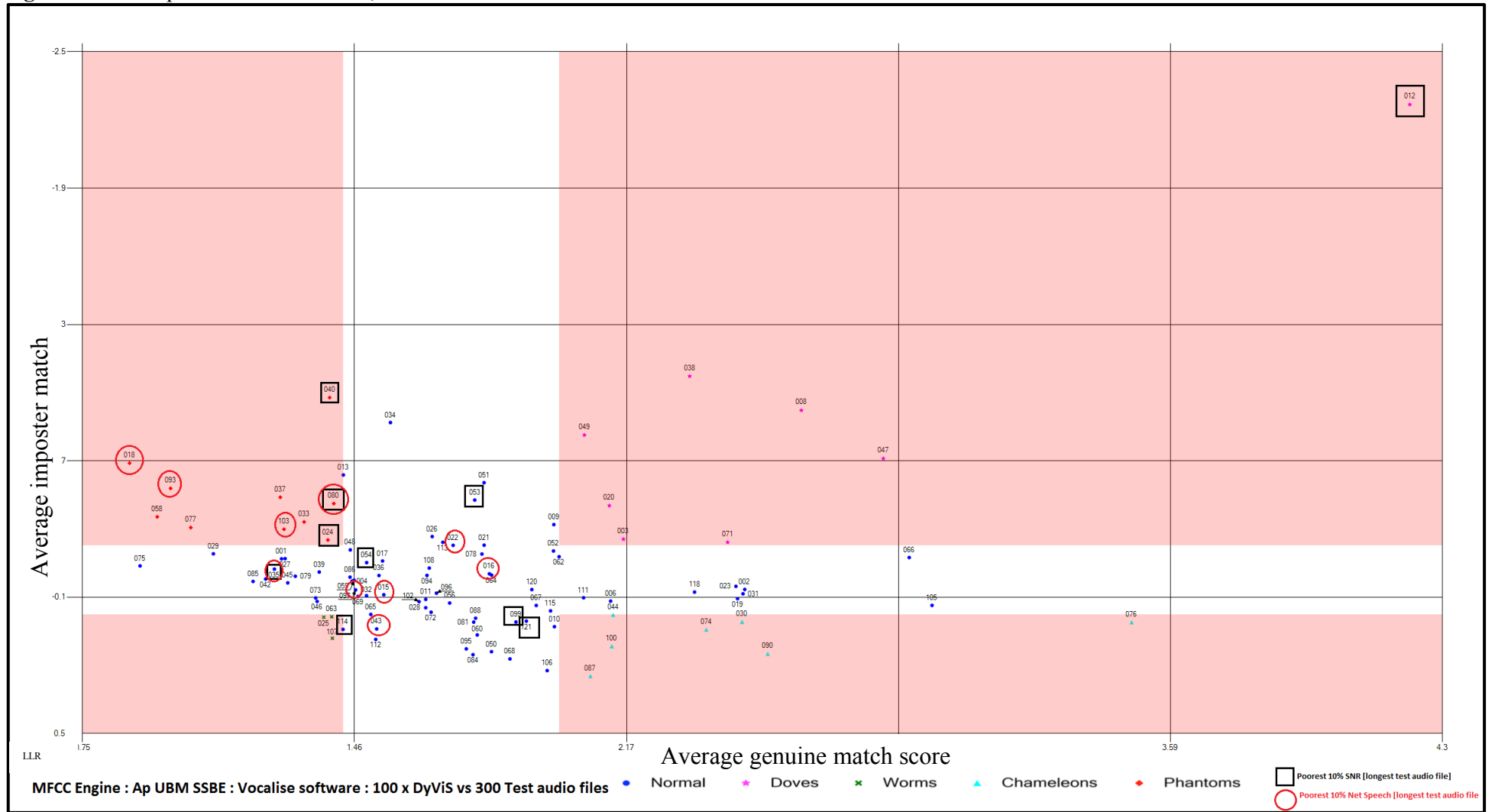
Average SNR was estimated for the speaker models using a state-of-the-art commercial application. This software was not used in later experiments due to insufficient documentation in terms of how SNR was calculated and an alternative was sought (see 3.5.6 re WADA). Nevertheless, results found that the speech files varied from between 17.76db to 40.56db SNR average. This suggested that microphone distance/gain was not likely to have been adjusted significantly (either manually or automatically) to compensate for speakers with differing vocal effort. Nevertheless, zoo plot examination of the files in the bottom 10% of the SNR range determined that there was likely to be a correlation with poor speaker performance and tendency towards left hand clustering (Figure 6.10 and Figure 6.11) with the exception of speaker 012. This experiment assisted with informing the

methodology for chapter 8, in terms of applying controlled degradation using the addition of noise and determining more accurate approach for measuring SNR.

### **6.5.7 Net Duration Test Results**

Preliminary examination of net duration for test audio file 3 (i.e. residual from editing the SM and first 2 TA files) provided the following zoo plots (Figure 6.10 and Figure 6.11). Whilst, again, zoo position was not conclusive in itself - this preliminary test suggested speakers with lower net duration may tend to appear towards the lower left-hand side of the zoo plot and this was further examined in the main net duration experiments completed in chapter 7.

Figure 6.10: Zoo plot 100 SM x 300 TA, MFCC GMM-UBM. Net duration and SNR tests







## 6.5.8 Zoo Plot Position

The position of any speaker on a zoo plot is naturally dependent on the other speakers in that test (i.e. relative). For example, removal of outliers created further outliers and a shift of region boundaries. Speakers previously positioned on classification boundaries moved from normal and were classified as animals (and vice versa). To examine this further in the preliminary tests, each of the animal groups was removed in turn and the baseline test re-run. It was found that this then created a statistical wave effect as each of the average imposter match scores adjusted. This was more notable in the y-axis values, due to the weighting of inter and intra speaker variation data. For example, in assessing inter-speaker variability, the data is rich as scores are calculated from multiple cross comparisons in this test 29,700 or  $(99 \times 3) \times 100$ . Data used to generate genuine match scores was constrained to just three genuine TA files per speaker in these tests.

In the baseline data, the genuine match data was also edited from single session data and it is important to be mindful of this. Whilst extracting multiple test audio from the same session has the advantage of reducing cross channel contamination, it would be preferable to more accurately capture intra-speaker variability through multiple non-contemporaneous sessions. These could better reflect the variation in speech likely from effects such as mood change or fatigue, for example. However, for the purposes of the main acoustic variability experiments conducted the same session data was preferable to limit the conflation of additional (session) variables.

## 6.6 Responses to Questions

The following are responses to the questions posed in 6.2.

**Q1: How should baseline be best established and what are the optimum ASR settings?**

**A1:** Various baseline tests were conducted with different ASR settings and performance (EER%) was measured. The prepared test data performed well and an optimum EER% was reached that was assessed to be consistent with a state-of-the-art MFCC GMM-UBM systems (1.244%).

**Q2: Are the zoo plot classifications of speakers similar for MFCC and LTFD?**

**A2:** There were some similarities in terms of speaker scores/results (8% of speakers appeared in the same zoo plot quadrant for both MFCC and LTFD extraction engines). However, as expected, the different methods of feature extraction produced variation with respect to zoo plot positions.

The LTFD system was likely to be more resilient to noise in some circumstances (re speaker 012) although the MFCC system performed better overall (EER%). In light of this, and because MFCC systems are much more widely deployed, the LTFD system was discontinued for the subsequent main experiments.

**Q3: Are zoo plot classifications affected by the technical features of the recordings, i.e. SNR and net speech duration, rather than just those features intrinsic to the voices themselves?**

**A3:** As demonstrated by the preliminary tests completed in reference to SNR and net duration positioning and by the behaviours of the speaker 012 files it is highly likely that performance is strongly influenced by the technical qualities of the recordings. Examination of voice quality produced some consistencies in terms of clustering/and general zoo plot position for speakers scoring high with lax larynx characteristics and those referred to as ‘breathy’. Further research is recommended but this tentatively demonstrated that other intrinsic factors also influence speaker performance in ASR systems and therefore zoo plot position. Risk was identified in terms of potentially conflating variables (VQ and acoustic variability) and aspects of high intrinsic variability, such as the initial vocal effort of speakers should be examined in the experiments pertaining to SNR.

## 6.7 Conclusion

The preliminary tests guided the scope and methodology of the subsequent experiments with the following recommendations.

- i. Highly controlled process(es) are required to artificially degrade the baseline data under measurable conditions which do not introduce additional variables.
- ii. Additional research into voice quality and intrinsic speaker variability for ASR is recommended. However, for present purposes it was decided that the main experiments should be confined to investigating acoustic variability where measurements can be obtained. Subjective perceptual data, whilst informed by experts, was not used.
- iii. The pace and continuing evolution of ASRs was such that several iterations of updates were introduced during the preliminary testing. Version control will be essential to ensure that any observations relating to performance are as a direct result of acoustic variability and not patch/version updates. Experiments should be adapted to incorporate modern i-vector systems.
- iv. It was shown that automatic LTFD analysis could potentially be more noise resistant than MFCC feature extraction (e.g. mains hum and speaker 012). On the surface, this could appear to offer benefits over MFCC. Nevertheless, the overall EER% was higher in the LTFD results than the MFCC showing poorer overall performance. Another potential option, of fusing results together from both types of systems, was stated in Jessen, Alexander and Forth (2014) and Gold, French and Harrison (2013). However, directly fusing MFCC and LTFD methods (i.e. LR scores) was later tested and found to provide no significant performance benefit (Hughes et al., 2017b).
- v. Inhibitors and contaminants affected different speakers in different ways. However, since baseline speaker scores and positions (zoo plots) were established and any additional

acoustic variability would influence those positions – it was determined that intrinsic variability should not be a limiting factor in proceeding to larger scale experiments examining ASR performance and broader acoustic influence.

# Chapter 7 Net Duration

---

## 7.1 Introduction

In conjunction with quality, the quantity of speech available for comparison is a key variable to be considered when using automatic speaker recognition systems (ASRs).

It is widely accepted that as net duration decreases ASR performance (EER%) deteriorates. However, it can be difficult to determine the quantity of speech required to achieve an acceptable level of ASR performance and confidence in outcome. In broad terms, net duration becomes more significant when comparing brief audio files (<1m) for both speaker model and/or test audio. The experiments conducted in this chapter therefore examine the influence of short net speech duration (<1m) on human assisted automatic speaker recognition systems in detail.

The chapter begins with a literature review to provide context. Three sets of experiments are then conducted. All experiments use the 100x male DyViS speaker data (task 1, mock police interviews). The speech files were edited to create 1m speaker models with two 1m test audio files per speaker of defined net duration. Baseline performance was established using both the OWR Vocalise (GMM-UBM) and OWR iVocalise (i-vector/PLDA) ASR systems.

In the first set of experiments 30 tests were completed (15 x GMM-UBM ASR and 15 x i-vector ASR) with net duration decreased for both the speaker models (SM) and test audio (TA) files. These were decreased at 5s iterative steps, with 1s steps from the sub 5s point. For experiment 1 the SM and TA files were of matched duration. Results were compared to baseline with the objective of broadly comparing the performance of two types of ASR systems (GMM-UBM and i-vector/PLDA) and determining how resilience to very low net duration compared. Metrics for equal error rate (EER%) and cost of likelihood ratio (Cllr) are presented.

In the second set of experiments both the speaker models and test audio files were reduced in 5s iterative steps with 1s steps below 5s and a full set of cross comparisons was completed at all durations for both SM and TA files – i.e. 1m SM compared to TA of 1m, 55s, 50s, 45s, 40s... then 55s SM compared to TA of 1m, 55s, 50s, 45s etc. This was completed using only the i-vector PLDA system due to file acceptance. The objective of this experiment was to provide a highly detailed analysis of performance with full metrics to examine potential thresholds for optimum performance and minimum net duration acceptance for a modern state-of-the-art system. Results are presented on 15 x 15 comparison grids for both EER% (overall ASR performance) and Cllr (accuracy).

The third set of experiments combined 50 x speakers with short duration (20s) test files and 50 x speakers with 1m x 2 test audio files. These were then compared against the 100 x 1m (baseline) speaker models using only the i-vector PLDA system. The objective of this experiment was to examine potential ASR performance risk (false accept rate, false reject rate) when combining different lengths of test files within the same set of comparisons.

Results are presented with discussion. Practical recommendations for casework and at-scale ASR integration are presented. The chapter concludes with suggestions for future areas of research.

## 7.2 Background

From experience, audio recordings are often inherently limited in nature and circumstance can sometimes preclude opportunities to obtain both a long, validated, speech sample(s) for the speaker model (SM) and/or questioned material (TA). Net duration can be influenced by many factors including channel dependency. For example, in applications such as push to talk radio communication (PTTR), utterances can have a tendency to be quite brief in nature and speech obtained for comparison/verification can often be as little as several seconds.

Since the early development of speaker recognition systems, applying ASRs to low net duration speech has been an enduring technical challenge. This gave rise to one of the initial research questions that motivated this chapter. Can the recent improvements in modern speaker recognition systems provide improved performance under very low net duration conditions or will the error rate always remain high? I.e. below a certain net duration threshold there won't be enough speech information to conduct an ASR comparison, but what is that point?

ASR manufacturers often claim that their latest system provides greater accuracy on shorter speech files. Yearly competitions are run by the National Institute of Standards and Technology called the Speaker Recognition Evaluations (NIST-SRE). At the competitions, the best performing systems are benchmarked using very low net duration speech from standard corpora (5s and 10s) such is the significance to the forensic speech community in progressing the technology. There is also an enduring requirement to better understand the performance of new extraction methods, statistical modelling algorithms and obtain representative metrics for performance on low duration speech. The rate of improvement is fast. ASR systems are continually evolving and the systems for pattern matching are becoming more sophisticated. Even within the timeframe for writing this thesis the progress of performance improvements has been observed with the commercial availability of i-vector systems. During the final months of writing this thesis, new x-vector and deep neural network approaches (Snyder et al., 2018) have been developed which effectively apply machine learning to further improve performance (Kelly et al., 2019).

Quality and quantity of speech in the context of ASR comparison is linked, with the former often influencing the latter. Examples include interference, intermittent noise, speaker to microphone proximity (i.e. movement), dropouts/faults, overlapping speech (with an interlocutor) or variable network bandwidth/transcoding. An uncooperative speaker can also influence net duration or where intra speaker variability is high and/or where modal voice is not used frequently enough within the submitted recording(s) such as shouting, screaming, whispering, out of breath or intoxicated etc.

In addition, it has also been noted from experience that intelligibility reasoning is often incorrectly applied to the anticipated reliability of ASR attribution. Even when a sufficiently large quantity of speech is presented for comparison it may be that only a very small fraction of the recording(s) is assessed as technically acceptable for ASR analysis and/or passes the speech detection phase.

To summarise, the central objectives for the experiments were:

- i. To measure the performance of a standard GMM-UBM ASR system and an i-vector PLDA ASR system under controlled conditions to complete a broad comparison on low duration speech performance;
- ii. To obtain comprehensive reference data for a state-of-the-art i-vector ASR system performance (EER%) and accuracy (Cllr) metrics to provide detailed information on operating and performance thresholds to assist with informing speech acceptance criteria for ASR use;
- iii. To examine the risks associated with combining short and long duration test/questioned audio within the same set of comparisons.

### **7.3 Additional Definition of Terms**

It is important to define the term ‘short net speech duration’. In an overview of research relating to net duration Poddar, Sahidullah and Saha (2015) stated:

“There is no standard definition of short duration in ASR. However, we observed that most of the published literature considered segments of duration 5-10 sec as short utterances for experimental evaluation and analysis.” Poddar, Sahidullah and Saha (2015: p.93)

The above definition is accepted for the purposes of this chapter.

In further defining terms it is important to state that net duration here applies more to the quantity of speech successfully passing the speech detection phase, rather than the quantity of speech as edited by a human prior to ASR analysis. This is because the speech passing the detection phase is invariably shorter. This can be due to the speech detection phase removing certain unvoiced, or

lower volume, speech sounds which fall below a certain threshold (i.e. perceived by the machine as silence when it is not). In addition, the speech detection algorithm completes further removal of between word silences (i.e. additional concatenation). The removal of multiple sections of silence and low amplitude speech therefore reduces net duration overall. For example, files edited carefully to 1m were notably reduced down to as low as 40s to 54s after passing through speech detection phase. On iVocalise, post-processed net duration values are extracted and so these values are also referenced in the experiments (net duration range).

Within class covariance normalisation (WCCN) is widely attributed to Hatch, Kajarekar and Stolcke (2006). In WCCN multiple speech samples, usually from different sessions and/or channels, from the same speaker are aggregated. This can create a richer set of speaker model data all assigned to the same speaker and ASR performance is improved from better separation of channel from speaker data. Whilst WCCN was applied to research systems that informed the experiments, it was noted that it did not significantly improve results. In addition, the experiments completed in this thesis are in a single channel domain. Finally, there was the potential that WCCN could add additional and unknown variability. WCCN is therefore referred in reference to the literature review but not applied to the experiments completed in this thesis.

## **7.4 Literature Review**

This section places the subsequent experiments conducted into context with specific regard to ASR performance and net speech duration research.

A very early study by Bricker and Pruzansky (1966) recorded short duration speech samples from 10 speakers and played them back to 16 listeners. All were known to each other at Bell Telephone Laboratories. The listeners were asked to match utterances to pictures of speakers. Their work confirmed research findings from Pollack, Pickett and Sumbly (1954), Clarke (1965) and Voiers (1961;1964) that, for humans at least, duration was linked to the accuracy of identification, whilst appreciating that other perceptual factors also contributed (Voiers, 1961;1964).

Bricker and Pruzansky (1966) demonstrated that the number of phonemes was linked to duration - i.e. intra speaker variability was constrained by constraining duration resulting in a loss of phonetic variation (i.e. speech data quantity and variety). This in turn provided lower accuracy scores from the listeners. Research such as Bricker and Pruzansky's (1966) study demonstrated that short duration has a negative influence on the human perception of speaker identity. Whilst humans rely on familiarity (and memory) and the subsequent experiments in this thesis focus solely on ASR systems and Bricker and Pruzansky (1966) offered a prophetic quote on the use of computers for speaker verification.



‘...we are in a position of wondering why the human needs information that the computer doesn't have in order to do as well.’ Bricker and Pruzansky (1966: p.1448).

Kanagasundaram et al. (2011) compared the performance of a joint factor analysis (JFA) (Kenny et al., 2006) i-vector ASR systems on 2008 NISTSRE data - results are recreated in Tables 7.1 and 7.2.

**Table 7.1:** Training and truncated test data (part I). Kanagasundaram et al. (2011: p.2344)

Utterance Length Training, or SM, to TA	i-vector JFA System EER%
2s to 2s	<b>35.25</b>
4s to 4s	<b>30.48</b>
8s to 8s	<b>23.39</b>
10s to 10s	<b>21.17</b>
20s to 20s	<b>12.79</b>
50s to 50s	<b>6.51</b>
2.5m to 2.5m	<b>3.37</b>

**Table 7.2:** Training and truncated test data (part II). Kanagasundaram et al. (2011: p.2344)

Utterance Length Training, or SM, to TA	i-vector JFA System EER%
2.5m to 2s	<b>22.48</b>
2.5m to 4s	<b>17.96</b>
2.5m to 8s	<b>13.43</b>
2.5m to 10s	<b>12.11</b>
2.5m to 20s	<b>7.67</b>
2.5m to 50s	<b>4.54</b>
2.5m to 2.5m	<b>3.37</b>

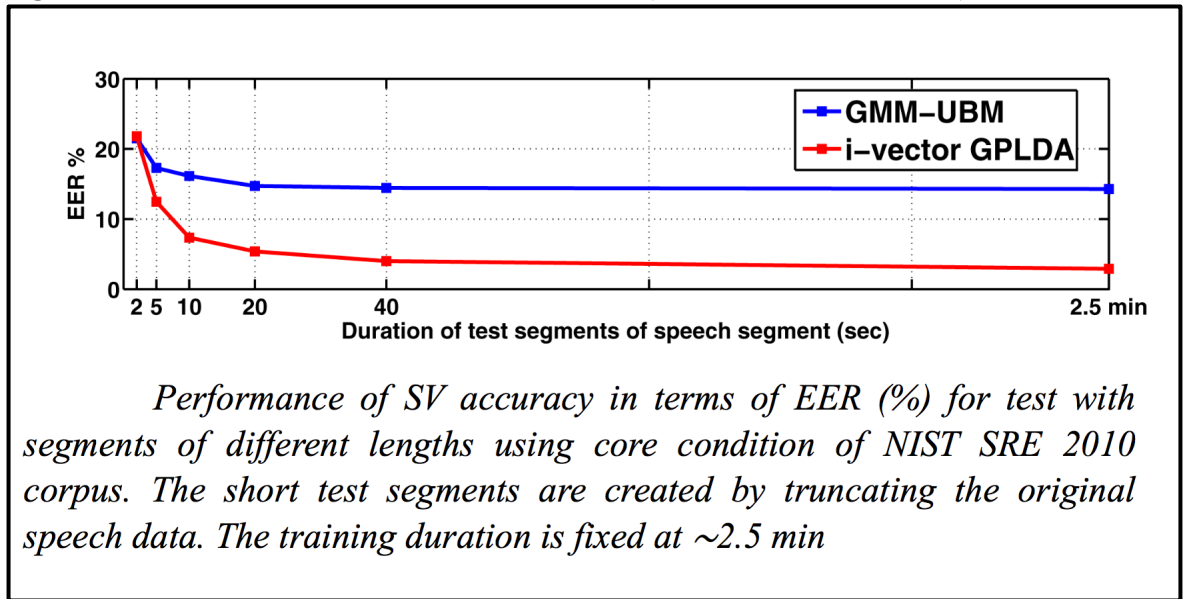
Their research demonstrated marginal improvements using alternative system architecture. Nevertheless, the benefit of longer training material (SM) was clearly evident with results indicating performance decline below 10s (10s speaker model to 10s test audio). Kanagasundaram et al. (2012) later developed a system for improving performance under short duration conditions by training the PLDA on short utterances (or S-Norm) on NIST (2004/5/6) SRE data.

Poddar, Sahidullah and Saha (2015) produced a comprehensive overview of research relating specifically to speaker verification and short utterances, including a useful summary of research completed, up to 2015, with corresponding error rates at the 10s SM or 10s TA net duration (see Appendix I). To summarise, results from the 10 different research studies showed high variability, with EER ranging from 21.56% to 4.29%. These results could be explained by many factors including feature extraction method, the different training conditions/NIST normative data and system settings. Nevertheless, their research summary provided several, broad observations that assisted with informing the research questions in this thesis. For example, training data (TA) over 10s improved equal error rates largely irrespective of other variables, raising the question - would this be the case on a much more modern i-vector system or would they exhibit greater tolerance at <10s? Finally, it was noted that WCCN was applied to multiple tests including both the best and worst performing results - suggesting that WCCN was likely to have a marginal influence on performance.

Larcher et al. (2014) studied the lack of phonetic variability as net duration decreases, using the ALIZE i-vector ASR toolkit on RSR2015 data (using male data only). The RSR dataset initially comprised of 300 English speakers (143 female and 157 male). The average duration of recordings was measured at 3.2 seconds. They demonstrated that EER% improved by same phrase pronunciation (SM and TA). In doing so, they confirmed research findings from others - citing Larcher et al. (2012) who also stated that the lexical content affects ASR performance. Larcher et al. (2013) applied VAD which effectively removes non-speech frames prior to the statistical modelling process and it is suggested it was likely to have influenced results. This is because different utterances at varying vocal effort could effectively cause more or less speech to pass through the VAD stage, dependent on threshold. Nonetheless, the phonetic content of the utterance, i.e. what the speaker says, is still valid - as lower duration generally produces less phonetic variability. Das, Jelil and Prasanna (2016) also found that constraining the spoken text in speaker model and test audio influences ASR performance. The large reduction of variability in speech utterances, as net duration decreases, is clearly an important factor and becomes more significant as speech data is restricted to very low duration - as also found in early studies by Boise, Hebert and Heck (2004) and in Hebert (2008). This research prompted questions as to zoo plot position for best and worst performing speakers under very short net duration conditions. Would there be anything noticeable in the spectrograms pertaining to the better performing speakers at very low net duration in comparison to the poorest?

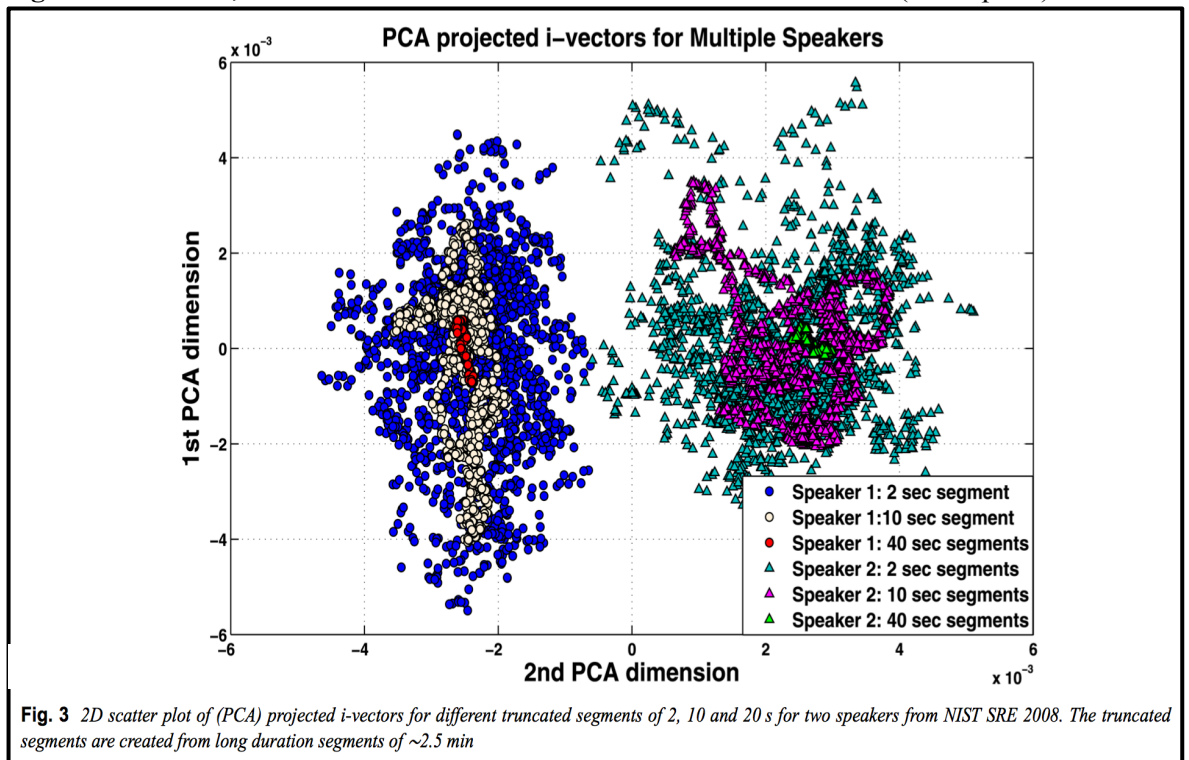
Poddar, Sahidullah and Saha (2015 and 2018) presented two graphs which also influenced the experiments in this chapter (Figure 7.3 and 7.4). The first shows the fall in EER% (i.e. better performance) for both GMM-UBM and i-vector/PLDA ASR systems on NIST SRE2010 data. Note the training material/speaker model was fixed (2.5m approximate).

**Figure 7.3:** Poddar, Sahidullah and Saha results 2015 (GMM and i-vector/PLDA)



The second graph (Figure 7.4) reproduces the uncertainty in (2D) i-vector point estimation as the net duration falls. Low net duration effectively disperses the vector space causing poorer discrimination.

**Figure 7.4:** Poddar, Sahidullah and Saha 2D i-vectors on low duration files (2018: p. 94)



Sarkar, Matrouf, Bousquet, and Bonastre (2012) examined the effect of early i-vector modelling on short and mismatched utterance duration. They used 2004 NIST SRE data to train (normative data) and 2008 NIST SRE data for speaker models and test audio. In 5s to 5s comparisons their modified i-vector system achieved 15.26% to 21.63% EER with 5.32% to 11.77% EER for 10s to 10s, dependent on training conditions. They concluded that a mixture of shorter and longer duration

training data was preferable when the questioned audio was brief. In circumstances where the duration of test audio comparisons was mixed, i.e. long and short, they concluded that longer training data was preferable. Somewhat counter to the research by Sarkar et al. (2012), Hasan, et al. (2013) noted that their early i-vector system trained on long duration utterances performed more poorly when presented with low duration questioned audio. Hasan, et al. (2013) team proposed three methods to compensate for mismatched duration; multi-duration PLDA training, score calibration and multi-duration PLDA training with synthesised short duration i-vectors. Overall, they found that the score calibration method was more encouraging in terms of compensating for duration mismatch, but they found that performance did not actually improve significantly for any of the methods suggested (comparative EER% figures were not provided). Nevertheless, for the experiments in this chapter the PLDA (session 1) was validated, in conjunction with OWR, to ensure the inclusion of low duration speech.

Fatima and Zheng (2012) coined the acronym SUSR (short utterance speaker recognition). They proposed that background noise becomes more influential as duration decreases. They also suggested that data segmentation is of greater importance at short duration since phoneme data could be lost if poor truncation occurs (i.e. at a higher percentage of the overall data) and it is widely known that (machine) speech detection and segmentation accuracy are an enduring weakness of the process. Fatima and Zheng (2012) also summarised by proposing six areas of research that could potentially improve performance in SUSR. Interestingly, these all related to combining speaker verification technology with prosodic mapping methodologies, rather than PLDA amendment or score calibration. Nevertheless, Chakroun, Frikha and Zouari (2018) supported this and have begun researching methods of potentially integrating additional speech information, for example from dialect detection, to improve SUSR.

The European Network of Forensic Science Institutes (ENFSI) published guidelines for the examination of speech for speaker verification (2015) and specifically provides recommendations on duration in regards to forensic semi/automatic speaker recognition which they refer to as FASR and FSASR.

‘Many FASR and FSASR methods require that the ‘net duration’ (i.e. pure speech from the relevant speaker, with all irrelevant information removed or disregarded) is no shorter than about 15-30 seconds. There is no general rule about the amount of audio material necessary and different methods might have different requirements. Ultimately, the minimum net duration required for a method has to be established with a method validation or other tests.’ ENFSI Section 5.4.1, (2015: p.33).

With the recent improvements in i-vector speaker verification systems this raised the question as to whether the minimum limit could be amended downwards?

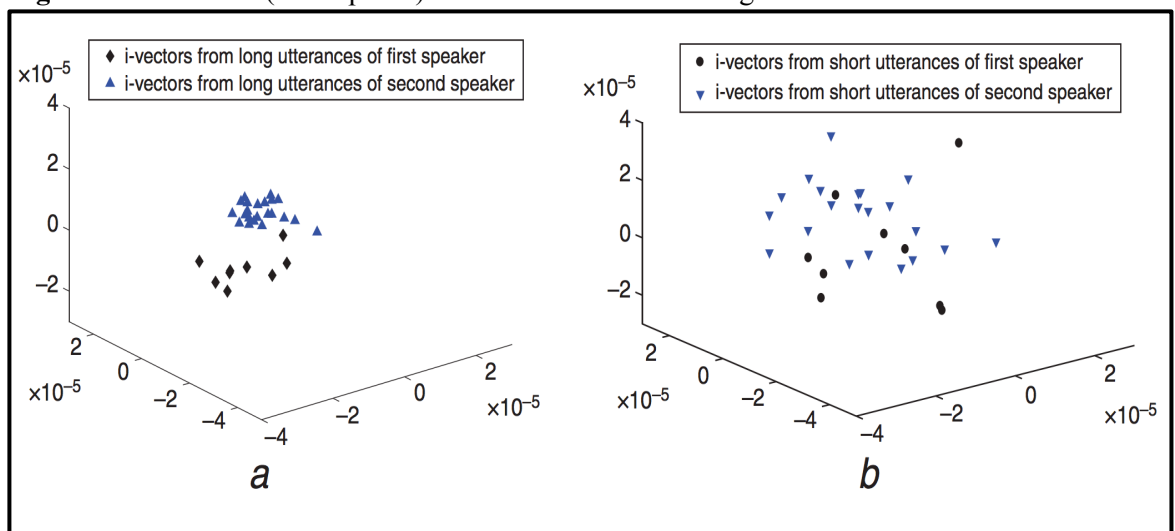
Bhattacharya, Alam and Kenny (2017) recently demonstrated that i-vector system performance can degrade when presented with very short duration recordings (<10s). In benchmarking their i-vector/PLDA system, primarily to test convolutional network performance (outside the scope of this chapter) the team used both the NIST, SRE 2010 test set and speech data from previous evaluations (NIST SRE 2004 to 2008) to generate speaker models and test audio files. A portion of NIST material was held back to create a bespoke normative set (PLDA). Their tests used 4,032 unique speakers from both genders. The SM and TA were edited to 10s and 5s respectively. The i-vector system used was not specified. The results from both 5s to 5s tests and 10s to 10s tests produced 24.78% EER and 17.44% EER. The team also observed that i-vector/PLDA systems appear vulnerable to performance issues (greater EER%) with extremely short audio (<5s). A view as to why this was so was not presented but a plausible suggestion is that it is due to the greater dispersal of i-vectors as found by Poddar, Sahidullah and Saha (2018, p. 94) shown in Figure 7.4 - and simply not enough speech data to create a robust/accurate enough statistical model.

Ma et al. (2017) supported Poddar, Sahidullah and Saha with their explanation as to why i-vector ASR systems do not produce significantly better performance over other types of ASR under short net duration conditions.

‘...due to limited phonetic coverage, statistics estimated from a short duration utterance are not as representative of the acoustic space as those from a long utterance. This then makes the distribution of i-vectors estimated from short utterances different from that of i-vectors from long utterances for the same speaker...’ Ma et al. (2017: p.405).

The group also illustrated the dispersal difference between long and short samples in 2D and highlighted the reduction of clustering (reproduced in Figure 7.5).

**Figure 7.5:** Ma et al. (2017: p.405) 2D i-vectors and short/long net duration



## 7.5 Questions and Hypotheses

From the research completed the following questions were raised. This section presents those questions with associated hypotheses.

**Q1: How does a state of the art i-vector/UBM, TV, LDA+PLDA system perform in comparison to a GMM-UBM system under low net duration speech conditions?**

**H1:** In reference to previous research and with improvements in statistical modelling the i-vector PLDA system should marginally outperform the GMM-UBM system with respect to both EER% (discrimination performance) and Cllr (accuracy). This should be more significant for very short duration utterances (under 15s) due to the improvements in statistical modelling in the i-vector/PLDA ASR system.

**Q2: For the i-vector system, is performance degradation linear or are there any identifiable tipping points? If so, what are the optimum net duration settings for performance and net duration acceptability?**

**H2:** Research by Bhattacharya, Alam and Kenny (2017) et al. demonstrated that i-vector ASR performance degraded as net speech duration fell below 10s for both SM and TA. It is expected that this will be broadly replicated. However, since the iVocalise system and underlying normative data are different to their research system, their performance figures will not be exactly reproduced.

**Q3: If 50 x speakers from the baseline test audio (i.e. 1m x 2 for 50 speakers) are compared with 50 speakers from the shorter duration test audio (i.e. 20s x2 per 50 speakers) is zoo plot position influenced by net duration when using 1m (baseline) speaker models for all 100 speakers?**

**H3:** It is suggested that the 50 speakers with shorter duration test audio files should cluster towards the lower left in the zoo plot. Conversely, the 50 x longer duration speakers should place towards the upper right, producing higher true positive/match scores and lower false positive/imposter scores. However, the duration of 20s was specifically chosen so as to narrow the differential between baseline and test conditions. It could therefore be argued that zoo plot positioning may not vary significantly enough to cause noticeable separation/clustering.

**Q4: In the very short duration results (e.g. 1-3s) is there any noticeable lexical/phoneme commonalities or spectrogram observations that explain zoo plot positioning for speakers who perform well (Doves)? Conversely, do the very poor performing speakers (Worms, Phantoms, Chameleons) exhibit high lexical divergence or any notable spectrogram observations?**

**H4:** In reference to previous research it is hypothesised that higher similarity between speaker model and test audio could improve speaker performance so this could be reflected in zoo plot position. However, it could also be argued that zoo plot position may be as a result of other or

conflated variables. In relation to spectrogram observations, and in line with previous research, it is likely that audio files for speakers who perform better at low net duration simply contain more speech information.

## 7.6 Methodology

As documented (5.3), the DyViS speech files were edited to remove silences and speech from the interlocutor (including overlapping speech). The audio files were cut to length using the Twisted Wave batch processing application [twistedwave.com/mac](http://twistedwave.com/mac). Output was dip sampled (approximately 10%) to validate that the application was accurate and did not add artefacts.

To establish baseline performance the control set was created. The first edited minute was used to create a speaker model (SM) for each of the 100 speakers. The subsequent 2 minutes generated x2 test audio (TA) files, per speaker, at 1m for each file. Residual speech was discarded in this chapter as, for some speakers, there was insufficient audio to generate a third 1m TA file.

Batch processing was then applied to an exact copy of the baseline data to generate each of the test data sets, constraining the net duration accordingly. Thirty test datasets were created for experiment 1 (15 x GMM-UBM and 15 x i-vector ASR comparisons). Net duration was decreased for both the speaker models (SM) and test audio (TA) files at 5s iterative steps with 1s steps <5s.

For experiment 1, the SM and TA files were matched in terms of duration. Results were compared to baseline. For experiment 2 both the speaker models and test audio files were reduced in 5s iterative steps with 1s steps below 5s. A full set of cross comparisons was then undertaken at all durations for both SM and TA files using the i-vector PLDA system. Experiment 3 combined short duration (20s) test files from 50 speakers and baseline test audio files (1m) from 50 different speakers to compare them against 100 baseline speaker models (1m). This experiment used the i-vector/PLDA system. The OWR ASR systems used are specified in 5.1.2, 6.4 (i.e. the SPOKE UBM for the GMM-UBM system) and Appendix G with the following adaptations.

- i. The threshold for minimum net duration acceptance was set to zero to prevent enrolment rejection.
- ii. Normative data for both systems did not include any of the DyViS dataset and the PLDA included both long and short duration speech files (as per previous research recommendations). The range of net duration passing the voice activity detection (VAD) stage was logged for each test to provide additional detail and, in all cases, the absolute VAD net duration was lower than the pre-VAD values.

Results were examined using OWR Bio-Metrics software version 1.8.0.704 (2017).

Note that attempts to recalibrate the system, to compensate for performance degradation, were consciously not taken to avoid conflating variables. In instances where the Cllr (accuracy) is above 1.0, but the EER% (discrimination) is low, this suggests that the ASR is operating relatively effectively.

## 7.7 Results

**Experiment 1.** The Table below (7.6) clearly shows the performance differential between the GMM-UBM Vocalise results (tests 1-15) and the iVocalise i-vector/PLDA results (tests 16-30). System performance is represented in equal error rate (EER%) and accuracy in cost of log likelihood ratio (Cllr). All files (100 speaker models and 200 test audio files) were accepted with the exception of test 15 where the duration was constrained to the point of files failing to pass enough audio to the statistical modelling phase. Therefore, for test 15, EER% and Cllr were calculated for ‘passed’ audio files only (194 SM and 19,206 TA elements).

**Table 7.6:** Net duration experiment 1, GMM-UBM and i-vector/PLDA results, matched

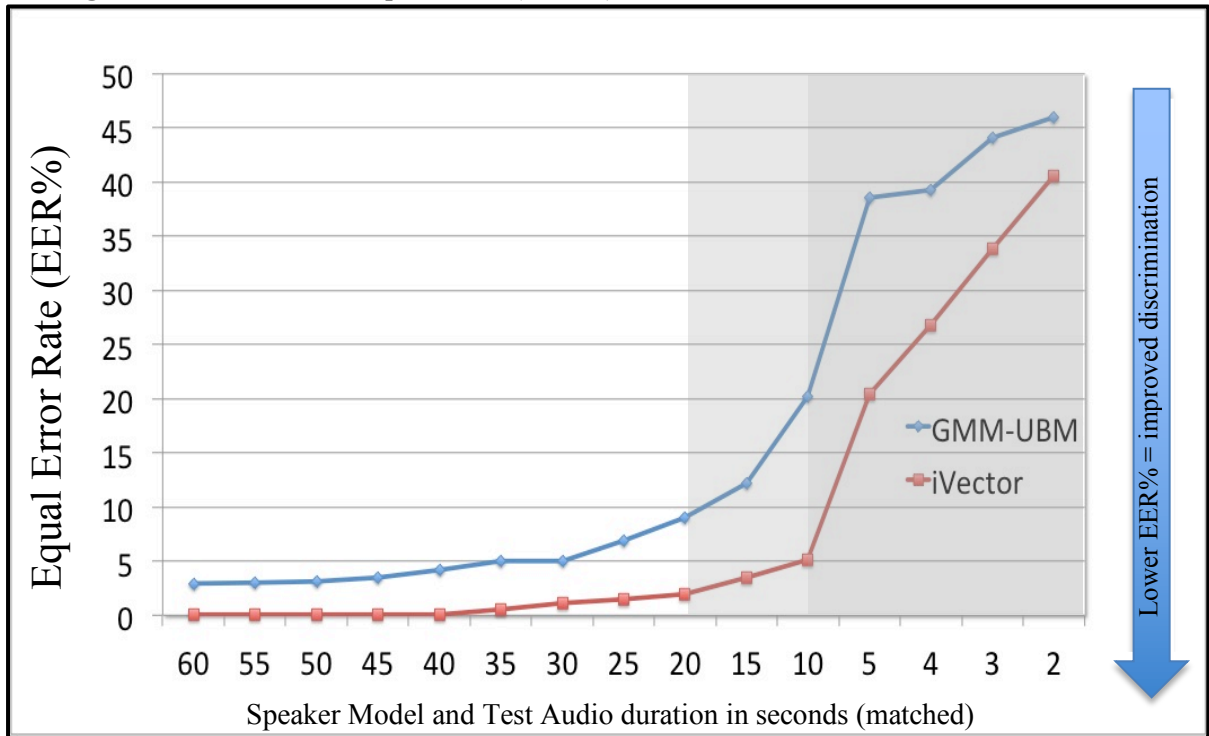
Test #	Speaker Model Duration (seconds)	Test Audio Duration X2 (seconds per file)	GMM-UBM EER%	GMM-UBM Cllr		Test #	i-vector EER%	i-vector Cllr
1	60	60	2.932	0.726		16	0.005	0.087
2	55	55	2.998	0.697		17	0.008	0.074
3	50	50	3.099	0.660		18	0.020	0.060
4	45	45	3.528	0.650		19	0.018	0.042
5	40	40	4.131	0.610		20	0.030	0.031
6	35	35	5.020	0.564		21	0.495	0.029
7	30	30	5.033	0.527		22	1.124	0.138
8	25	25	6.881	0.516		23	1.477	0.381
9	20	20	9.033	0.566		24	1.995	1.138
10	15	15	12.212	0.810		25	3.495	3.832
11	10	10	20.144	1.723		26	5.149	11.793
12	5	5	38.505	5.096		27	20.391	35.198
13	4	4	39.263	5.971		28	26.798	42.058
14	3	3	44.084	6.042		29	33.866	49.660
15*	2	2	45.993	5.471		30	40.549	58.526

\* Only 194 SM and 19,206 TA elements passed VAD

Figure 7.7 and Figure 7.8 show the decline in performance re net duration for both systems (EER% and Cllr).



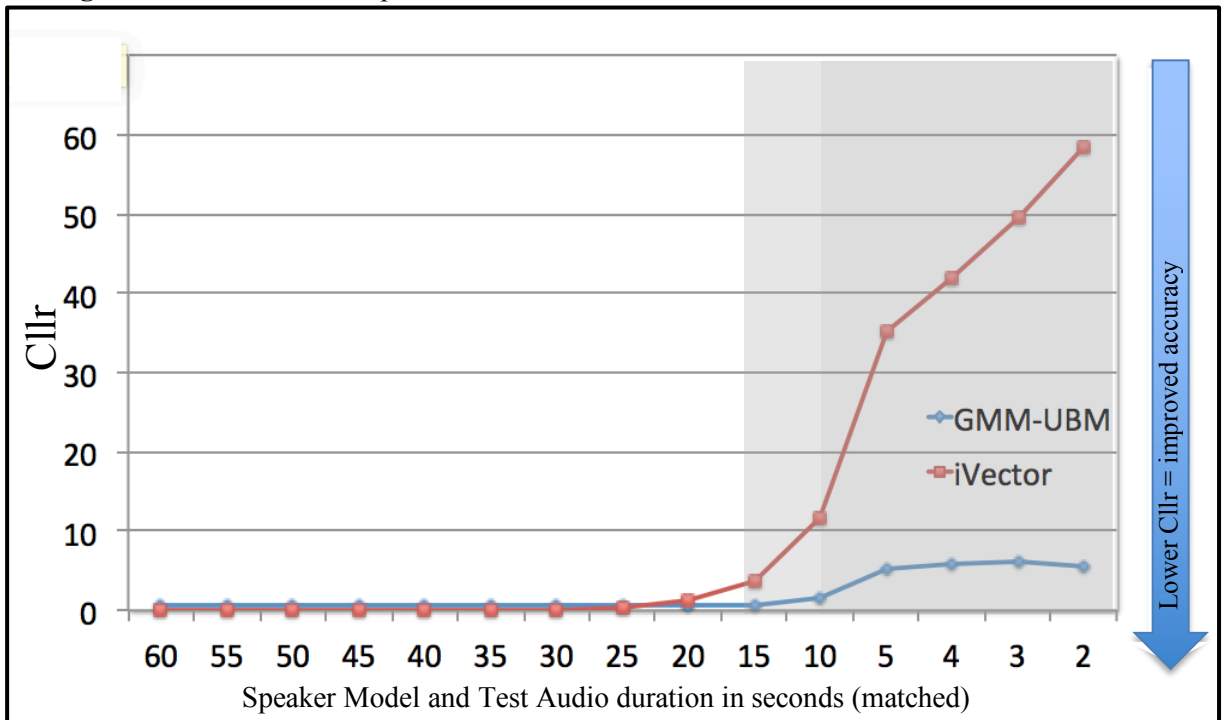
**Figure 7.7:** Net duration experiment 1(a and b). EER%, i-vector and GMM-UBM



■ Performance tipping points identified for both systems.

Greater resilience to low net duration conditions shown in i-vector results (discrimination and accuracy, Figure 7.8).

**Figure 7.8:** Net duration experiment 1a and b. Cllr, i-vector and GMM-UBM, matched



**Experiment 2.** Tables 7.9 and Table 7.10 (next 2 pages) document the full EER% and Cllr results for all 225 cross comparison tests completed using the i-vector/PLDA ASR system.

**Table 7.9:** Net duration experiment 2. EER% Results. IVocalise, i-vector system

EER% Results		Test Audio (seconds)														
		60	55	50	45	40	35	30	25	20	15	10	5	4	3	2
SM Seconds	VAD Pass (net range)	40 to 54	47 to 50	33 to 45	29 to 40	25 to 35	22 to 31	18 to 26	15 to 22	12 to 17	8 to 13	6 to 9	1 to 4	1 to 3	0.5 to 2	0.5 to 1
60	40 to 52	0.005	0.005	0.005	0.005	0.013	0.015	0.020	0.106	0.573	0.995	0.992	4.503	6.492	10.000	17.035
55	36 to 48	0.005	0.008	0.013	0.008	0.015	0.035	0.053	0.498	0.917	1.033	1.124	4.874	7.013	10.444	17.014
50	32 to 44	0.010	0.008	0.020	0.018	0.025	0.078	0.121	0.518	1.479	1.515	2.018	4.864	7.588	11.018	16.360
45	30 to 39	0.008	0.010	0.023	0.018	0.025	0.078	0.442	0.897	1.513	1.492	1.977	4.578	8.528	11.397	18.078
40	26 to 35	0.008	0.013	0.020	0.020	0.030	0.111	0.500	0.990	1.487	1.510	2.005	4.510	7.487	11.902	18.063
35	22 to 31	0.379	0.379	0.063	0.061	0.109	0.495	0.871	1.000	1.495	1.498	2.025	5.301	7.957	12.010	18.573
30	19 to 26	0.078	0.399	0.379	0.429	0.498	0.540	1.124	1.412	1.505	1.985	2.518	4.869	7.518	12.485	20.243
25	16 to 21	0.419	0.452	0.765	0.462	0.500	0.929	1.510	1.477	1.525	2.490	2.871	5.487	8.134	11.919	21.131
20	12 to 17	0.498	0.558	0.593	0.609	0.912	1.492	1.498	2.407	1.995	3.005	3.886	7.076	9.078	14.490	21.555
15	9 to 13	1.003	1.010	1.005	1.434	1.498	1.934	1.990	2.505	2.989	3.495	3.957	8.457	12.588	17.184	22.552
10	5 to 8	1.503	1.604	1.884	1.692	1.990	1.998	3.025	3.490	4.559	5.520	5.149	10.995	15.490	19.619	25.174
5	2 to 4	7.970	8.472	8.982	9.523	9.503	9.523	9.033	10.533	10.886	10.477	11.543	20.391	23.773	27.482	32.328
4	2 to 3	9.960	10.480	11.874	12.505	12.096	11.477	12.543	12.657	12.525	13.998	15.018	23.841	26.798	27.957	33.737
3	1 to 2	13.487	15.402	15.553	15.111	15.033	15.843	16.351	16.472	17.866	19.025	18.823	25.753	30.487	33.866	38.617
2	0.5 to 1	23.225	24.538	25.379	23.886	23.823	23.634	24.788	24.399	25.179	26.912	27.058	33.099	36.553	38.205	40.549

Colour is indicative of relative performance and does not denote acceptability criteria or threshold(s).

**Table 7.10:** Net duration experiment 2. Cllr Results. IVocalise, i-vector system

Cllr Results		Test Audio (seconds)														
		60	55	50	45	40	35	30	25	20	15	10	5	4	3	2
SM Seconds	VAD Pass (net range)	40 to 54	47 to 50	33 to 45	29 to 40	25 to 35	22 to 31	18 to 26	15 to 22	12 to 17	8 to 13	6 to 9	1 to 4	1 to 3	0.5 to 2	0.5 to 1
60	40 to 52	0.087	0.081	0.073	0.065	0.056	0.048	0.039	0.027	0.067	0.222	0.401	8.098	14.376	23.529	37.862
55	36 to 48	0.080	0.074	0.067	0.060	0.051	0.044	0.036	0.031	0.122	0.300	0.535	8.678	15.008	24.163	38.510
50	32 to 44	0.071	0.066	0.060	0.054	0.046	0.038	0.031	0.040	0.183	0.362	0.688	9.374	15.665	24.694	38.620
45	30 to 39	0.056	0.052	0.047	0.042	0.036	0.030	0.026	0.059	0.217	0.398	0.869	10.104	16.525	25.478	39.129
40	26 to 35	0.046	0.043	0.039	0.035	0.031	0.025	0.033	0.091	0.251	0.426	1.071	10.755	17.200	26.193	39.591
35	22 to 31	0.040	0.037	0.034	0.031	0.027	0.029	0.072	0.151	0.304	0.584	1.567	11.559	18.031	26.681	39.903
30	19 to 26	0.033	0.030	0.027	0.025	0.029	0.061	0.138	0.232	0.399	0.755	2.040	12.446	19.007	27.471	40.385
25	16 to 21	0.022	0.022	0.033	0.026	0.051	0.116	0.216	0.381	0.611	1.162	2.652	13.926	20.366	28.347	40.800
20	12 to 17	0.062	0.086	0.128	0.135	0.160	0.274	0.467	0.746	1.138	2.010	4.173	16.215	22.481	30.043	41.835
15	9 to 13	0.284	0.385	0.513	0.552	0.570	0.805	1.216	1.863	2.517	3.832	6.534	19.257	25.470	32.155	42.970
10	5 to 8	2.126	2.359	2.621	2.701	2.867	3.394	3.941	4.934	5.985	8.208	11.793	24.897	30.250	35.794	45.119
5	2 to 4	14.371	14.870	15.231	15.489	15.997	16.687	17.339	18.489	19.831	22.236	25.065	35.198	39.019	43.300	50.258
4	2 to 3	20.920	21.466	21.730	22.005	22.572	23.134	23.801	24.701	25.868	27.927	29.919	39.052	42.058	45.883	51.556
3	1 to 2	30.855	31.235	31.361	31.618	31.899	32.477	32.914	33.318	33.973	35.597	36.932	44.060	46.605	49.660	54.457
2	0.5 to 1	43.712	44.030	44.022	44.154	44.252	44.553	44.860	45.025	45.388	46.189	46.397	51.255	53.104	55.236	58.526

Colour is indicative of relative performance and does not denote acceptability criteria or threshold(s).

Cllr scores <1 are generally considered acceptable, scores >1 indicate a system with low accuracy (Hughes et al., 2019).

**Experiment 3.** Figures 7.11 and 7.12 present 2 zoo plots generated from the combined short (20s) and longer duration (1m) TA files compared to 1m SMs. Note the clear left/right clustering.

**Figure 7.11:** Experiment 3. IVocalise 50 speakers, 1m SM, 2x 1m TA comparisons circled

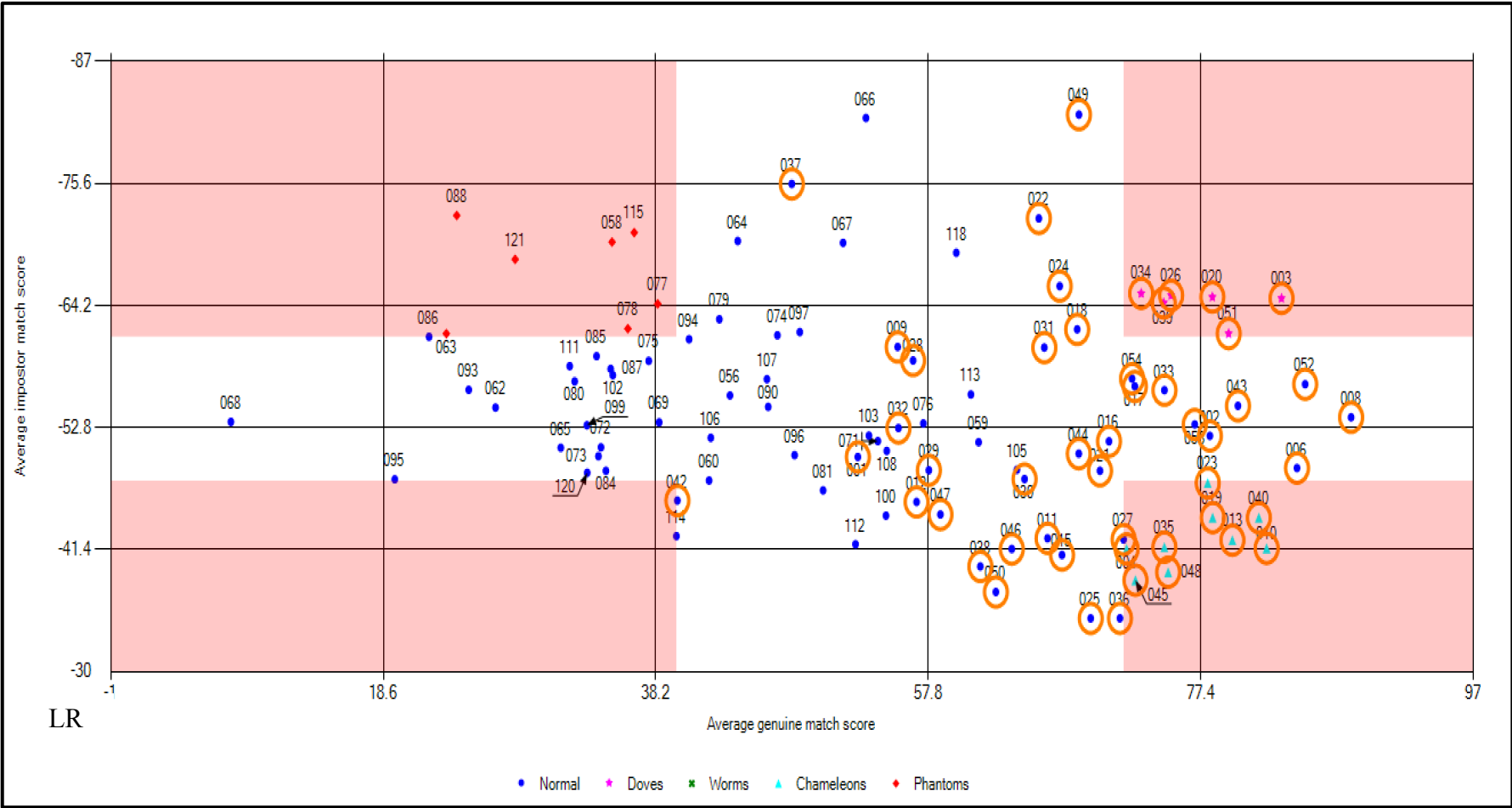
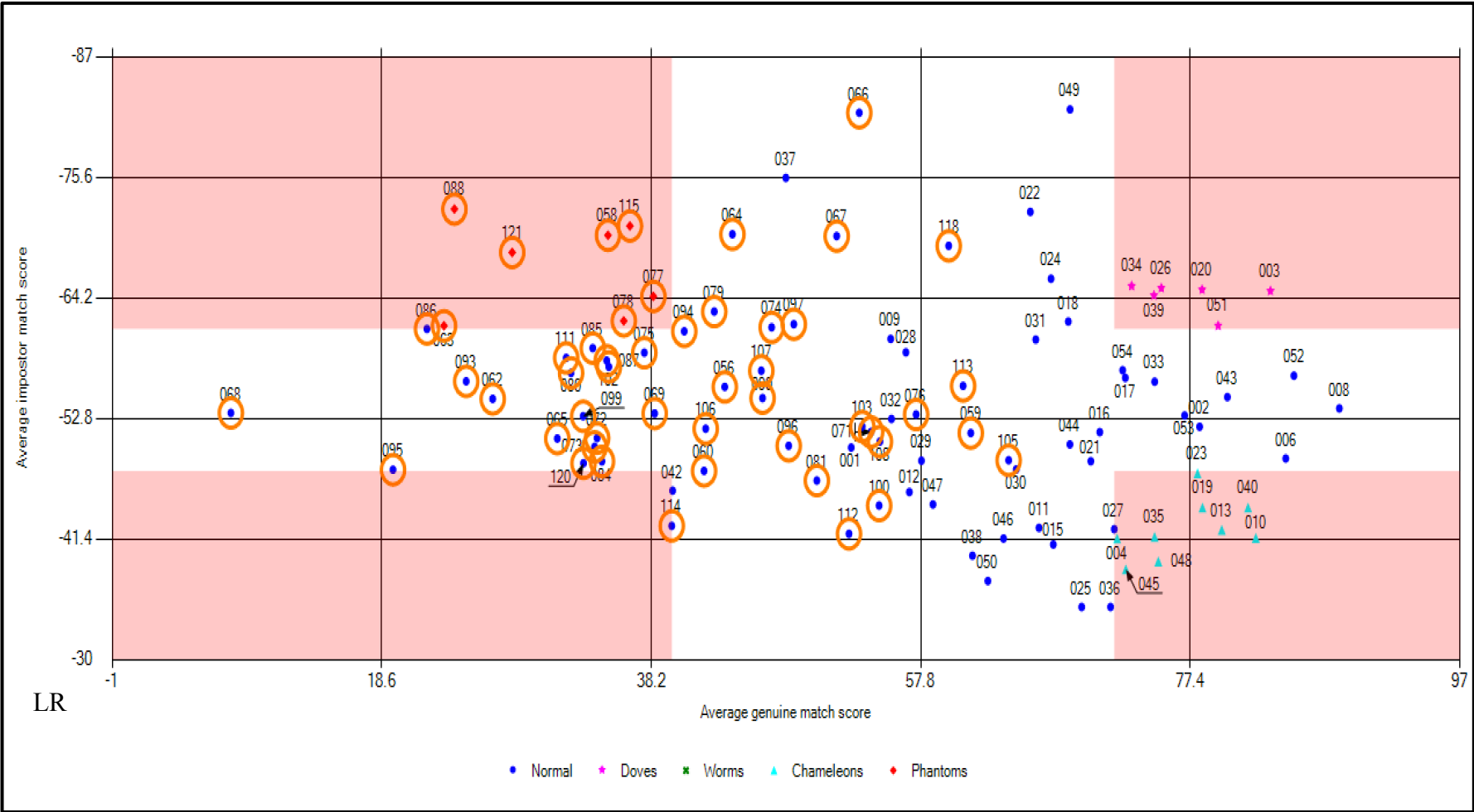


Figure 7.12: Experiment 3, IVocalise 50 speakers, 1m SM, 2x 20s TA comparisons circled



## 7.8 Responses to Research Questions

**Q1: How does a state of the art i-vector/UBM, TV, LDA+PLDA system perform in comparison to a GMM-UBM system under low net duration speech conditions?**

**A1:** As hypothesised the i-vector/PLDA system outperformed the GMM-UBM system at all test durations in experiment 1 for EER%.

At the lowest net duration setting (0.5s to 1s for both SM and TA) the EER% for both systems initially appear to be broadly similar (45.99% for GMM-UBM and 40.55% for i-vector PLDA). However, not all comparisons passed the VAD in the GMM-UBM test and so the EER% result is based on less data in comparison to the i-vector system and the underlying normative data is different between the systems. Nevertheless, results were also broadly consistent with previous (and recent) research, with some marginal improvements noted. In summary an i-vector system is expected to outperform GMM-UBM, assuming correct set up/normative data etc., likely due to improvements in feature extraction and statistical modelling density.

**Q2: For the i-vector system, is performance degradation linear or are there any identifiable tipping points? If so, what are the optimum net duration settings for performance and net duration acceptability?**

**A2:** Performance degradation was not linear. As predicted the i-vector system was more resilient to performance degradation at lower net duration and demonstrated a more gradual, shallower decline in EER% until the 10s (tipping) point. At the 10s point performance degraded sharply (effectively doubling in EER%). Tables 7.9 and 7.10 show this performance tipping point clearly. These results were consistent with research by Bhattacharya, Alam and Kenny et al. (2017) and are likely to be a result of poorer i-vector clustering, under very short net duration conditions, caused by a fundamental lack of speech information and low intra-speaker variability across the speech sample(s) in comparison with longer duration files.

For net durations of lower than 10s, performance degraded when the duration was constrained to the speaker model or the test audio and greater performance degradation was noted when both were reduced. Despite symmetrical (SM and TA) scoring – also discussed in 6.3 - this supports the view that there is a point where a lack of test audio data simply cannot be compensated for by using longer speaker models (or vice versa).

**Q3: If 50 x speakers from the baseline test audio (i.e. 1m x 2 for 50 speakers) are compared with 50 speakers from the shorter duration test audio (i.e. 20s x2 per 50 speakers) is zoo plot**

**position influenced by net duration when using 1m (baseline) speaker models for all 100 speakers?**

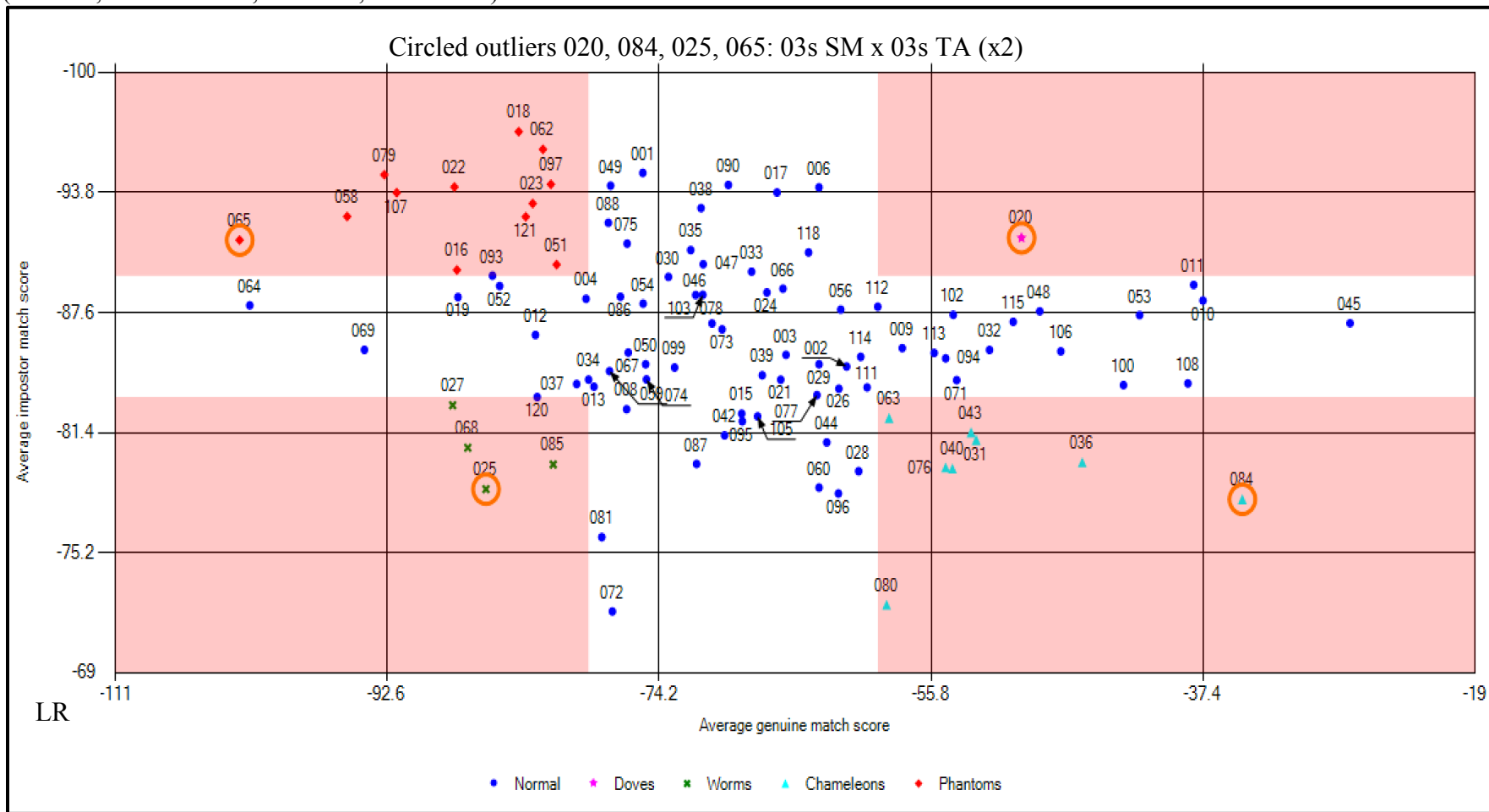
**A3:** Re the i-vector ASR system, whilst some speakers from the shorter duration comparisons performed well producing relatively high match scores and low imposter scores, there was a noticeable separation of results on the zoo plots (Fig 7.11 and 7.12). Those speakers with shorter duration audio files (2 x 20s TA) clustered to the left, with lower genuine match scores whilst longer duration speakers (2 x 1m TA) clustered to the right, with higher genuine match scores. It was noted that the average imposter scores appeared less affected (vertical plane of the zoo plot). EER% performance of the 20s TA files (0.573%) compared relatively favourably to baseline results (0.005%) results. Nonetheless, experiment 3 highlights the potential risk in combining low duration files with longer duration files within the same comparison set. I.e. lower match scores are likely to be obtained for short duration comparisons and high(er) match scores for long duration, potentially making threshold setting/score separation for variable audio lengths problematic (e.g. different net durations for suspect/genuine and imposter files could skew overall ASR results).

These results could also influence speaker model management – and it would be recommended that minimum and maximum net duration criteria are set to prevent uneven LR/LLR output (per speaker).

**Q4: In the very short duration results (e.g. 1-3s) is there any noticeable lexical/phoneme commonalities or spectrogram observations that explain zoo plot positioning for speakers who perform well (Doves)? Conversely, do the very poor performing speakers (Worms, Phantoms, Chameleons) exhibit high lexical divergence or any notable spectrogram observations?**

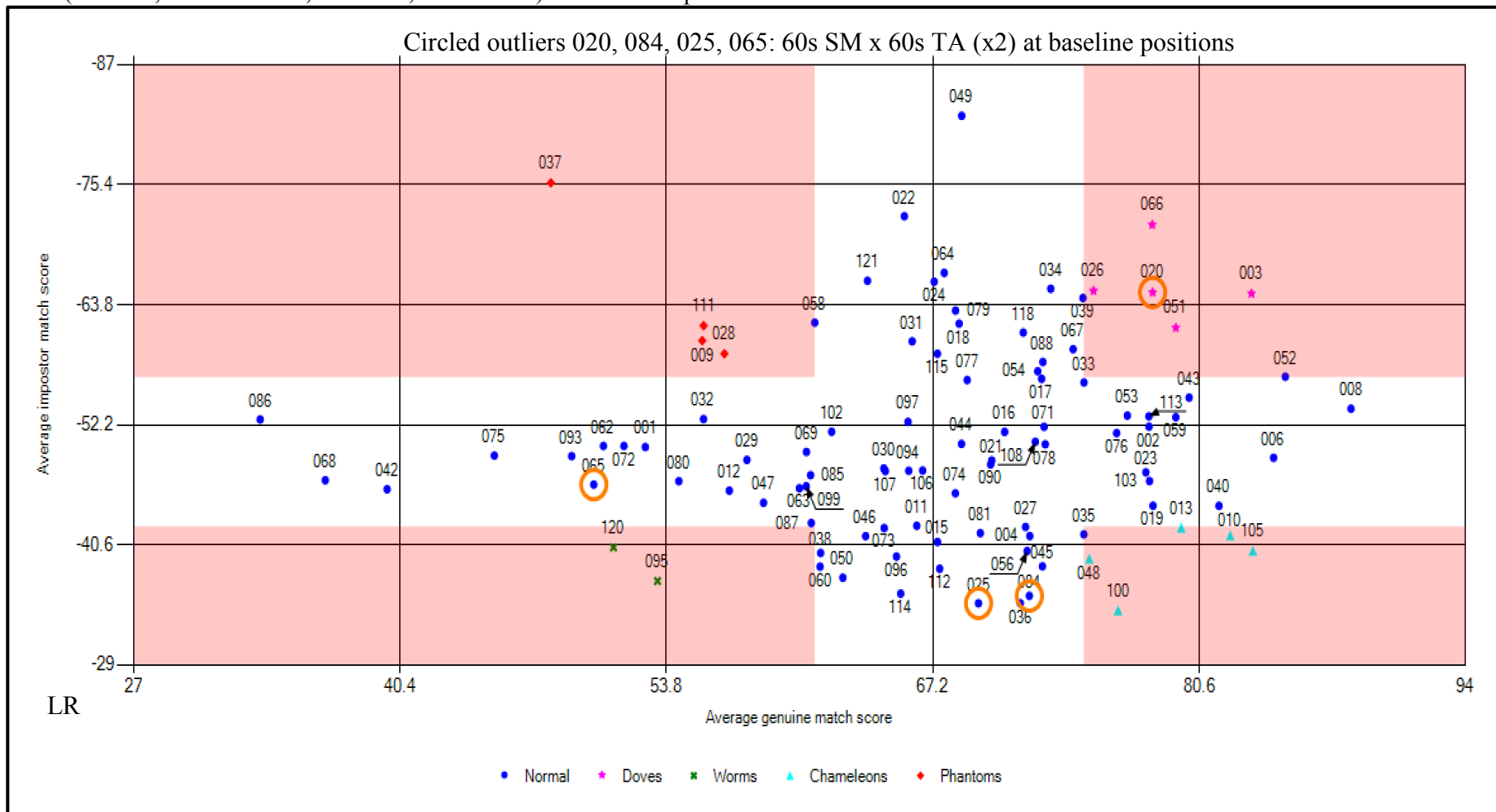
**A4:** On examination of outlier speakers within the 3s test results, e.g. Dove 020, Chameleon 084, Worm 025 and Phantom 065, no immediate correlation in terms of phonetic content and position could be identified (Figures 7.13 and 7.14).

**Figure 7.13:** iVocalise results at 3s SM x 3s TA (x2). 100 speakers, 4 outlier speakers circled (1x Dove, 1x Chameleon, 1x Worm, 1x Phantom)



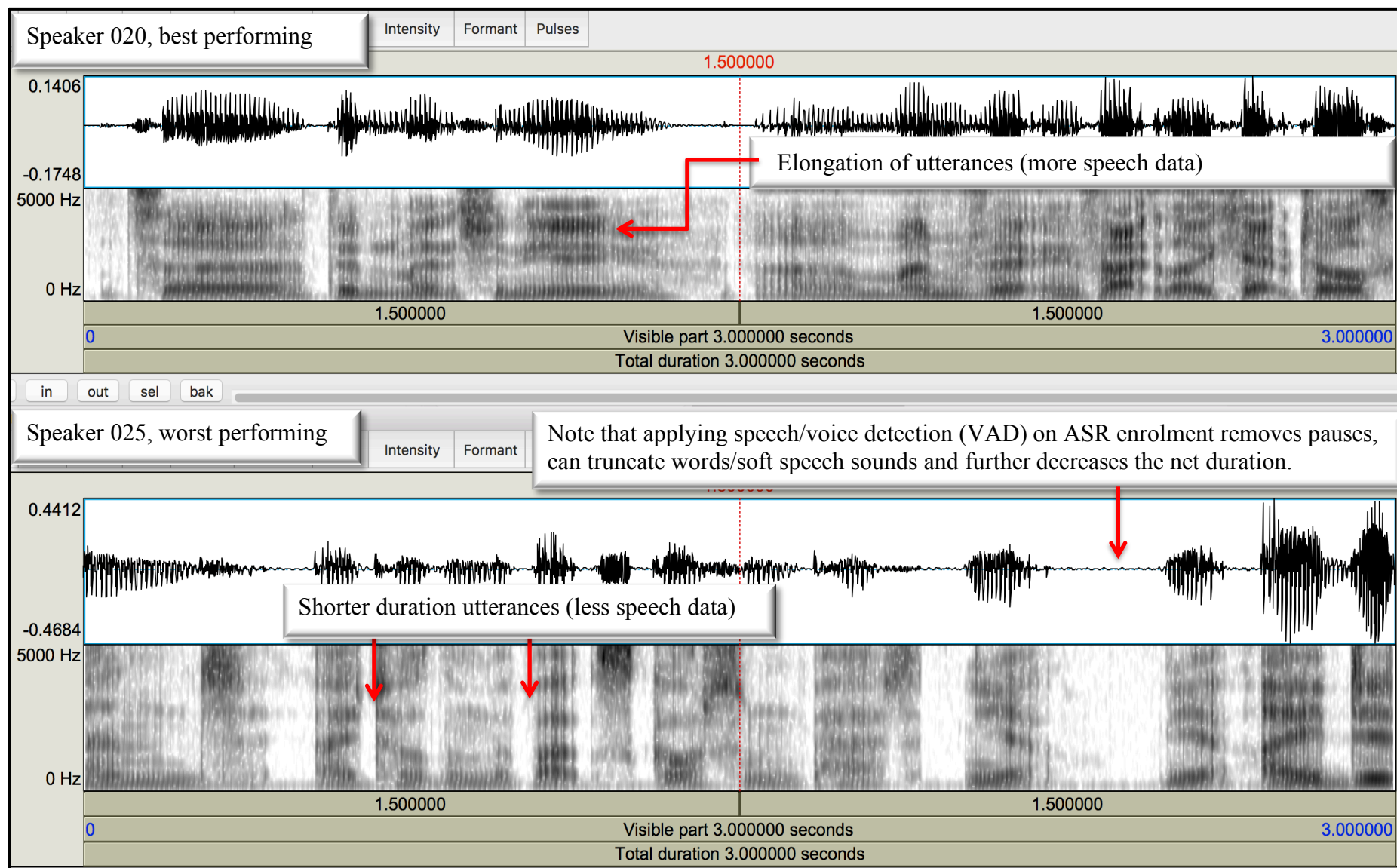


**Figure 7.14:** iVocalise results at 60s SM x 60s TA (x2). 100 speakers, 4 outliers as in 7.13. (1x Dove, 1x Chameleon, 1x Worm, 1x Phantom) - note shift in position



A full phonetic analysis was not completed as the utterances were so brief as to provide almost no useful speech data. Using spectrogram analysis (Figure 7.15) to examine speaker 020 (best performing) and 025 (worst performing), at 3s TA, there was a notable elongation of speech data (formants) for the top performing speaker in comparison to the poorest performing – who tended to use shorter utterances. Simply put, there was more speech data. In a long speech sample, this is likely less of an issue but in a short sample it is suggested that richness of data becomes more important. However, this hypothesis is somewhat inconclusive given the extremely small data sample. It is therefore suggested that, whilst phonetic information could potentially contribute to zoo plot position and poorer performance (as per previous research), other variables are likely to be conflated. So, whilst logic and previous research suggests that phonetic content could be a contributory factor, further research is required.

Figure 7.15: Praat spectrogram view of speakers 020 (best) and 025 (worst) 3s



## 7.9 Discussion and Practical Recommendations

Results from experiments showed marginal improvements in tolerance to very low net duration but were also broadly consistent with research outcomes from other ASR similar systems, as documented in the literature review. Throughout the experiments, both ASR systems consistently produced lower match scores and higher imposter scores for shorter net duration comparisons (<40s) and were of lower accuracy (higher Cllr).

For both ASR systems performance tipping points were found where performance severely degraded. For the more modern i-vector/PLDA system this became evident at the sub 10s net duration range where the rise in EER% was appreciable (5.149% compared to 0.005% at 1m, baseline). However, at the sub 5s band the EER rose considerably (20.391%) and so a non-linear decline in performance was observed. This clearly has implications with respect to ASR comparisons and speaker model management with low net duration continuing to have a negative influence on ASR performance despite improvements in system architecture, feature extraction methods, statistical modelling (i-vectors) and the use of larger scale & bespoke normative data (UBM, TV, LDA+PLDA). These experiments also further support ENFSI recommendations for minimum net duration thresholds (ENFSI 5.4.1).

It is important to note the influence that speech detection/VAD had on further reducing net duration (more so than human editing alone) and, in all instances, the human edited speech files were longer in proportionate terms than the post-VAD files. In terms of practical application, it would therefore be strongly recommended that post-VAD measurements (i.e. ASR file import reports) for net file duration should be documented and factored into the analysis when applying ASR systems. With respect to reporting, it can also be more difficult to determine the expected performance range (EER% and Cllr) for the ASR system itself and this must be reflected re confidence in output interpretation.

Whilst not directly conclusive from zoo plot positioning alone it is suggested that the lack of phonetic variation in extremely short duration samples (<1-10s) could be influencing performance at very low net duration. However, as phonetic variation was not explicitly examined, and auditory analysis could not be completed, a correlation to speaker performance could not be established and further research in this area is recommended.

It is hoped that the experiments have produced useful metrics, although these are offered only as a rough guide. Also, the extrapolation of thresholds from experiment 2 should not directly inform threshold(s) for different ASR system. Both previous research and the experiments completed support the view that performance/output can differ across ASR systems in respect to the normative

data, settings, calibration and audio quality – all of which can influence the ASR’s performance and accuracy on low net duration speech comparison.

In terms of practical application, results support testing and establishing system specific settings, i.e. minimum acceptable duration threshold(s) to mitigate against poor performance (EER% and Cllr). In addition, experiment 3 supported that comparative tests which more evenly apply net duration limits across both suspect and imposter files could assist with mitigating against skewing ASR results. Experiments suggest that thresholds would need to be carefully established on an ASR system so as not to exacerbate false positives/false negatives<sup>‡‡</sup> on low net duration comparisons.

The results from experiment 1 demonstrated that, in relation to Cllr (accuracy) an i-vector system may produce less accurate results under very low duration conditions than a GMM-UBM on similar length audio files, although it is argued that this is offset by much lower overall EER% performance on the i-vector/PLDA ASR. This is consistent with results from Poddar, Sahidullah and Saha (2018), who also demonstrated a fall in accuracy (Cllr) for i-vector ASR systems on very low duration speech files. A plausible explanation for this is likely due to the more precise clustering for the statistical modelling in the i-vector system (i.e. greater specificity). Interestingly, Poddar, Sahidullah and Saha (2018) also found that EER% discrimination performance began to decline below 40s with a similar tipping point located at approximately <10s/<5s. Experiment 1 results therefore support these findings with additional data.

Results from experiments 1 and 2 also demonstrated the risk in assuming GMM-UBM and i-vector ASRs provide similar performance and accuracy (EER% and Cllr) as duration declines. In practical terms this simply supports upgrading an ASR system, appreciating that a new or upgraded system should also undergo adequate performance testing, calibration and any net duration threshold adjustment(s) are based on objective testing (in relation to ENFSI guidelines) rather than manufacturers recommendations or previous ASR version settings.

If an ASR system is operated through an application program interface (API) – i.e. at command line level - comparison queries could be completed which are not as constrained by a more visible net duration threshold, more easily set and reviewed by an operator via a graphical user interface (GUI). Having net duration acceptance setting somewhat out of sight could be an additional risk factor, effectively enabling the bypassing of any recommended threshold(s) to force extremely short duration speaker comparison (<10s). Experiments show that that this would not be recommended

---

<sup>‡‡</sup> Low net duration audio events may still contain valuable information for an investigation.

and supports the view that ASR operators should ensure speech detection and segmentation processes are correctly configured and net duration thresholds carefully observed – with the documentation of post VAD duration.

Finally, further research is recommended to examine longer (than 1m) net duration comparisons to test if ASR performance can be further improved (e.g. 5m or 10m comparisons combined with 1m). It is plausible that a maximum performance saturation point will be reached with regard to statistical modelling and this may have been reached during experimentation. Nevertheless, research on long net duration could have implications in terms of better optimising speaker models and assessing how long they need to be, since they can be a significant resource cost with regard to (human) editing and management.

# Chapter 8 Signal to Noise Ratio

---

This chapter examines the effect of Signal to Noise Ratio (SNR) on ASR performance. In line with the overall objectives for the thesis the motivations for the experiments are:

- i. The production of metrics and reference material to assist with informing casework analysis;
- ii. To provide guidance on SNR thresholds for audio acceptance into ASR systems.

The chapter begins with a literature review of relevant research to provide context. Baseline performance is established for 100 x DyViS speakers (task 1, mock police interview data) on an OWR iVocalise i-vector ASR system using a bespoke normative set (UBM, TV, LDA+PLDA). To generate new test speech files noise was added to the baseline data to effectively decrease the SNR. Two different types of noise (white and pink) were applied at 10 iterative steps 5db apart (-45db to 0db). ASR tests were then re-run for both matched conditions (similarly degraded speaker model and test audio) and non-matched conditions (non-degraded SM and degraded test audio).

The GMM-UBM system was initially assessed, but could not be used effectively in the experiments conducted in this chapter. In early tests, the rejection of audio for a significant portion of the more heavily degraded data was observed despite multiple adjustments to speech detection thresholds and settings in an attempt to mitigate. The i-vector/PLDA ASR is therefore used throughout.

Further tests were also run to apply modern adaptive noise reduction techniques and normalisation to the baseline data in an attempt to positively influence the SNR and raise ASR performance.

Results and findings are presented. The chapter concludes with a discussion of the influence of SNR and acoustic variability on ASR performance. Practical recommendations are made to assist with informing speaker comparison under poor SNR conditions and the chapter concludes with recommendations for further areas of research.

## 8.1 Background

It is widely accepted that recordings with low noise relative to the speech signal are fundamental to accurate speaker comparison using ASRs. Estimation of SNR can therefore assist with providing metrics and define the terms ‘low’ or ‘high’ noise’ in the context of ASR performance. The confidence with which an ASR assessment is then made can also be better defined or, if the SNR is particularly low, decisions can be taken as to whether ASR analysis should be conducted at all with speech files rejected at the point of technical assessment. It was noted that the ENFSI

guidelines for best practice (Drygajlo et al., 2015) broadly reference ‘reduced SNR’ (2015: p.33) but do not specify db acceptance levels.

As previously discussed (3.2), SNR can be influenced by many variables at different points of the end-to-end signal chain. These include, but are not limited to, the performance of the microphone, the bit depth and sample rate, microphone proximity/vocal effort, the quality of the recording device (e.g. faults/susceptibility to interference) and environmental noise. SNR can also vary from moment to moment within an audio/speech event. Since many variables determine SNR, measuring and establishing the influence on ASR performance can only really be extrapolated from the use of controlled experiments, which do not directly replicate casework conditions. However, if it is possible to quantify the controlled conditions under which ASR performance deteriorates as SNR falls, then it should be possible to better predict how an ASR will perform under casework conditions.

## 8.2 Literature Review

Togneri and Pullella (2011) evaluated SNR variability on a GMM-UBM system with the addition of white noise on 64 speakers from the TIMIT database (630 x speakers, non-degraded/studio quality). Their experiments introduced white noise at 5db, 10db, 20db and 30db. They applied the G.712 codec and MIRS (Modified Impulse Response System) to simulate different channel characteristics. Cepstral Mean Normalisation (CMN) was applied, a method for removing the effect of non-speech from the cepstral values at the feature extraction stage. Note that CMN is similar to Cepstral Mean Subtraction (CMS) (Furui, 1981) which is integrated, by default, into OWR iVocalise. Even under relatively mild degradation of SNR Togneri and Pullella demonstrated GMM-UBM ASR performance declined (Table 8.1 from Togneri and Pullella, 2011: p. 37). Note that results are expressed in terms of percentage of correct comparisons, rather than EER%.

**Table 8.1:** Influence of SNR on GMM ASR system. Togneri and Pullella (2011: p.37)

<b>Effect of additive white noise (at different SNRs) and matched/mismatched channels on the 128 mixture GMM-UBM system and the standard features (with CMN).</b>			
	<b>G.712</b>	Results in % Accuracy	<b>MIRS</b>
<b>clean</b>	<b>94.5</b>		<b>94.5</b>
<b>30 dB</b>	<b>74.2</b>		<b>75.8</b>
<b>20 dB</b>	<b>42.2</b>		<b>39.8</b>
<b>10 dB</b>	<b>10.9</b>		<b>7.8</b>
<b>5 dB</b>	<b>3.1</b>		<b>2.3</b>



Evans et al. (2002) researched SNR in the context of the landline telephony domain. The database they used consisted of 2,000 speakers with 1,000 speakers used for model training (normative data/UBM). They demonstrated that adding 15db of car noise to the test audio from 1,000 speakers produced a drop in ASR performance on a GMM-UBM system (3 to 5 EER% compared to 36 EER%).

Research undertaken by Nakasone (2003) found that ASR performance on a GMM-UBM system began to degrade at approximately 16db SNR with a significant drop noted in score distributions at <14db SNR. Nakasone also noted increasing overlaps in LR plots showing true and false distributions effectively drawing together and merging at around 0db, which demonstrated the increasing difficulties encountered in casework when setting thresholds for poorer SNR comparisons. Nakasone's research supported the hypothesis that severely degraded audio (low SNR) should be regarded as unsuitable for ASR analysis and (independently) assessed for auditory analysis suitability.

Nakasone's research was further developed in Harmse, Beck and Nakasone (2006), which examined SNR and net speech duration to seek compensation algorithms. Their research comprised of 8 experiments under matched conditions using a bespoke corpus of fifty male speakers. The group encountered a common issue in the energy detection phase (determining speech over noise) which progressively lost accuracy as SNR decreased. This effectively produced a reduction in speech passing the ingest process, hence providing an additional motivation for their examining the link between SNR and net speech duration. However, assessing both SNR and net duration simultaneously raises issues with respect to isolating variables. Many experiments had very short and mixed duration speech samples (0.5s to 16s) which, as demonstrated in chapter 7 (net duration), can influence ASR EER%. Nakasone (2003) also previously demonstrated that using less than sixteen seconds of speech (either for test audio or speaker model) degraded ASR performance (3.9% EER for 16s x 16s (baseline) down to 50.8% EER for 0.5s x 0.5s). Nevertheless, broad tendencies were demonstrated that decreasing SNR produced corresponding poorer EER% performance. The research also produced a useful regression model for score compensation. In summary, the Nakasone research assisted with informing preliminary tests conducted in this chapter, particularly where it was determined that the speech detection phase required some adjustment simply to allow enough net speech to pass speech detection for the very low SNR speech samples.

Athulya, Vinashankar and Sathidevi (2017) encountered similar EER% performance effects on a GMM-UBM system with speech degraded under several conditions. They experimented with an alternative method of mitigating noise on speech data from the NOIZEUS corpus containing 30 IEEE sentences for 6 speakers (3 x males and 3 x females). They proposed the use of Gamma tone

filter cepstral coefficients (GFCC), which essentially model the way that the human cochlear works using overlapping band pass filters, as opposed to a standard MFCC feature extraction method (see chapter 3.4.3). The group also suggested using speech detection/VAD to spectrally subtract non-speech noise estimated values from speech and proposed a varying threshold scale in the VAD calculations. This they proposed would assist with better determining speech against noise. Their results are reproduced below (Figure 8.2).

**Figure 8.2:** Results from Athulya, Vinashankar and Sathidevi (2017: p.5)

TABLE II EER WITHOUT VAD AND SPECTRAL SUBTRACTION [19].			
SNR (dB)	EER in % (MFCC)	EER in % (LPC)	EER in % (GFCC)
0	40.48	40.48	30.48
5	38.57	38.1	28.1
10	38.1	28.57	28.1
15	35.71	28.1	24.29

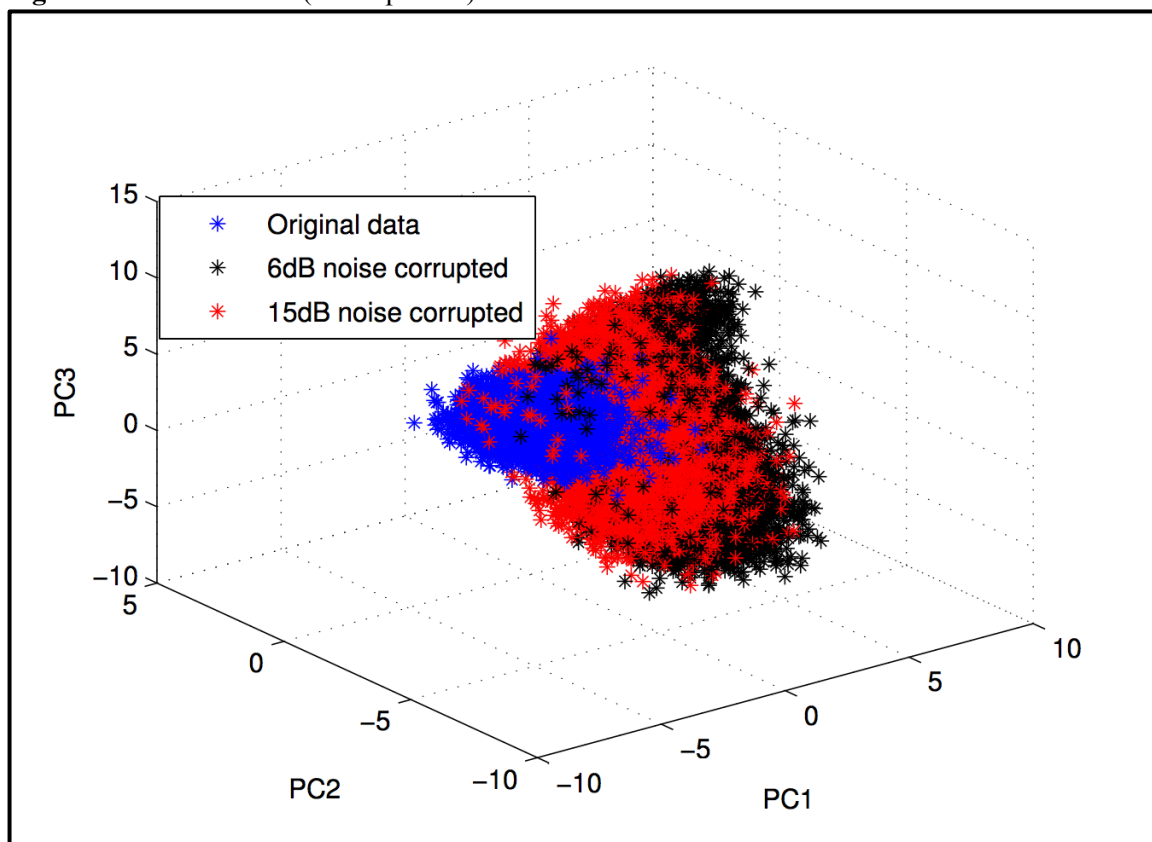
TABLE III EER WITH VAD AND SPECTRAL SUBTRACTION.			
SNR (dB)	EER in % (MFCC)	EER in % (LPC)	EER in % (GFCC)
0	38.57	37.62	23.81
5	38.1	35.71	21.43
10	35.71	25.71	19.05
15	28.57	23.81	16.67

Whilst GFCC and VAD approaches were shown to be beneficial to performance, in comparison with MFCCs, they conceded that the dataset was relatively small. In addition, some of speech samples that were held back were used in the UBM (universal background model or normative data). From the preliminary experiments completed, this was shown to skew results and produce artificially elevated performance in systems (3.5.6, chapter 6 and Appendix D). This is more noticeable if there are relatively small quantities of normative data and the addition of the test corpora is then a large(r) percentage of the overall. Nevertheless, the performance improvements were encouraging and this paper demonstrated a very innovative way, using different feature extraction methods, to improve ASR processing of low SNR speech.

Li and Mak (2015; 2016) demonstrated that utterances with similar SNR clustered together in i-vector subspace and, conversely, those with degraded SNR grouped apart. Their research was based on 7,156 utterances from NIST 20015-2008 SRE degraded with (speech) babble at 6db and

15db. They suggested that this observed shift could form the basis of performance improvements through the provision of bespoke normative set(s) (PLDA) to better accommodate variation in SNR.

**Figure 8.3:** Li and Mak (2016: p.5566) shift of mean i-vectors with SNR reduction



This hypothesis is similar to the concept of within class covariance normalisation (WCCN) where a speaker is effectively enrolled in multiple environments (in this case varying SNR) to inform the system that the speaker model is the same person with any i-vector variation predominantly caused by channel difference, in this case SNR. Note that other/different i-vectors could correspond to language, codec, frequency bandwidth etc. In the context of the experiments conducted in this thesis it was determined that manipulation of the normative data could produce an additional variable. The normative data session (UBM, TV, LDA +PLDA) was therefore fixed to maintain a constant as the SNR degraded. In addition, multiple models per speaker (i.e. WCCN) were not created, largely due to insufficient quantities of data (many sessions).

Beritelli (2008) examined the influence of background noise on SNR estimation in the context of speaker recognition. He experimented with 13 noise categories, examining the influence on F1, F2 and F3. Results showed that background noise has a varying influence on different formants and therefore vowel realisations. Beritelli (2008) recommended further work to examine SNR estimation at a sub-band level and this was further explored in Beritelli, Casale, Grasso and Spadaccini (2010). Their work completed a performance evaluation of several SNR measurement methods (manual, semi-automatic and 'real') highlighting some of the difficulties in under (or over) estimating SNR for speaker comparison and speech analysis. Their research examined the

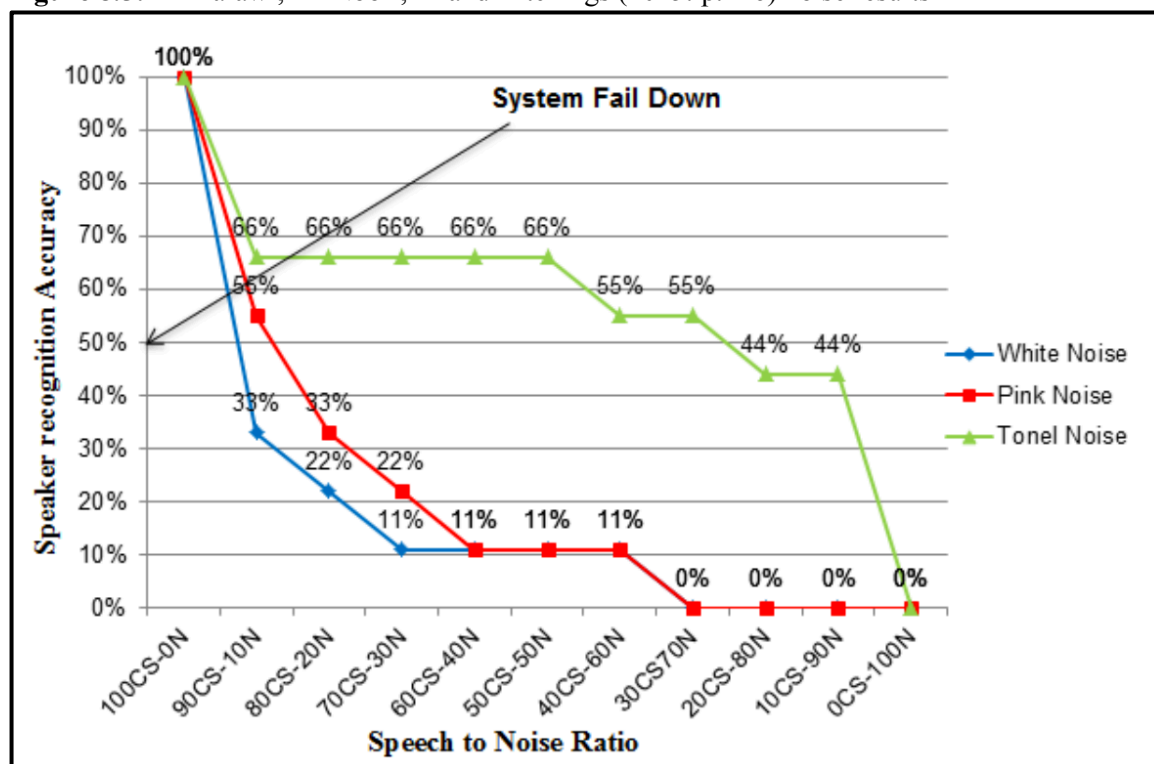
influence that noise (vehicle, office, crowds and construction) had on individual vowels (from TIMIT) and particular sensitivity was found for the diphthong /ai/. This led to the group requesting more effective SNR estimation algorithms. They also provided recommendations regarding the introduction of critical SNR thresholds for different speech sounds, although did not define them and it is suggested that this would be complex to implement.

Al-Karawi, Al-Noori, Li and Ritchings (2015) completed research experiments on the influence of noise (and, independently, reverberation - see chapter 9). The group used a Microsoft Speaker Recognition (MSR) tool kit to examine ASR performance. The toolkit can use either GMM-UBM or i-vector but for their research they selected only GMM-UBM. The team recorded 19 speakers (11 males and 8 females between 25 and 40 years old) at 16kHz sample rate. The speech samples were timed at between 30 and 40 seconds. The speech samples collected for the noise tests were text independent but also recorded in a different language. It is not clear if this conflated variables and it was also unclear as to the description of tonal noise. Their results are presented in % accuracy rather than EER% (Figure 8.5) so baseline (i.e. no noise at all – highest SNR) provided 100% accuracy. The term ‘system fail down’ was not fully explained (Figure 8.5) - but it is inferred that this meant equivalence to chance level accuracy. Normative data is not referred to. The poor performance of the MSR system was noted and the group recommended further investigation.

**Table 8.4:** Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) noise settings

SPEECH AND NOISE MIXING PERCENTAGE										
Speech	100	90	80	70	60	50	40	30	20	10
Noise	0	10	20	30	40	50	60	70	80	90

**Figure 8.5:** Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) noise results



Prasanna and Pradhan (2011) proposed that the VLR elements of speech are louder and have a higher SNR and are therefore likely to be more resilient to noise and poorer SNR recordings. They experimented with extracting the vowel-like regions, or VLRs, (vowels, semi-vowels and diphthongs) from speech using TIMIT and NIST 2003 corpora which they artificially degraded using NOISEX-92 data to demonstrate an overall improvement in EER% from 18.6% to 12.7% and 15.3% to 13.4%.

### 8.2.1 Vocal Effort and Signal to Noise Ratio

Speaking against environmental noise tends to cause the elevation of vocal effort. This is known as the Lombard effect, named after Etienne Lombard (1911), who studied voice elevation in the context of the hard of hearing and loud background speech. In the experiments completed in this thesis - artificial noise was added post recording and the Lombard effect was therefore not a variable. It is conceded that a speaker raising their voice could, broadly speaking, partially restore the SNR and therefore ASR performance. However, this would also introduce another variable as the increase in vocal effort would deviate from modal voice. Goldenberg, Cohen and Shallom (2006) confirmed that Lombard effected speech degraded ASR performance (2006: p. 237). On a GMM-UBM test system (2006: p. 233) they found EER% degraded by 10.1% overall (from EER 3.8% to EER 13.9%). They also noted that performance could be (partially) restored by transforming the Lombard speech by increasing the feature order – EER% 22.3% to 8.4% (p.237).

Jessen and Becker (2010), and Kirchhübel (2009) studied F2 and F3 values for 31 speakers. They reported that the variability between Lombard and normal speech was inconsistent across speakers and relatively small overall. In addition, the Lombard effect can introduce other consequences such as elevated first formant values and modification of voice - e.g. fundamental frequency and voice quality such as spectral tilt (Summers et al., 1988; Castellanos et al., 1996; Lau, 2008; Jessen, Köster and Gfroerer, 2005). Kelly and Hansen (2016b) also studied the specific influence of Lombard on ASR's – finding performance degraded (i-vector system). In summary, the influence of the Lombard effect and associated rises vocal effort were regarded as undesirable variables for this set of acoustic experiments – and, as stated, are not a feature of DyViS - but should clearly be considered in case examination.

Noise in audio recordings is often inconsistent, varying in a combination of intensity, duration and frequency content. SNR measurements, particularly in the context of the experiments conducted, are therefore estimates. To extract the estimated SNR, from the SM and TA files, the audio quality application Juicer (OWR, 2016) was used to provide consistency and batch analysis. Juicer extracts various metrics relating to the technical quality of audio including Waveform Amplitude Distribution Analysis, or WADA SNR estimate (see 8.2.2). This application and algorithm were assessed as providing less variability and bias than other methods tested. This then enabled more detailed analysis of results, for example in accommodating for the natural variation in vocal effort between speakers and the technical quality of the recordings.

## **8.2.2 Signal to Noise Ratio Estimation**

Kim and Stern (2008) developed Waveform Amplitude Distribution Analysis or WADA estimation and this was used for estimating SNR in the experiments conducted in this chapter.

Essentially, WADA SNR estimation uses statistical information calculated from the amplitude distribution of the speech waveform. This process is based on the assumption that (relatively) good quality speech has a Gamma distribution whilst background noise tends to have a Gaussian distribution. Kim and Stern (2008) concede that background speech or babble can also have a waveform distribution closer to Gamma in nature, but for the purposes of the experiments in this chapter (i.e. the addition of non-babble noise and a lack of background speech) the measurement was considered a valid form of SNR estimation.

The preliminary tests (Chapter 6) demonstrated that it was also important to recognise the strong connection between vocal effort and recording SNR (i.e. lower effort = lower signal). Speakers who talk quietly are likely to produce a lower speech signal in relation to either the background noise and/or require adjustment to the microphone gain level (upwards) at the signal input. This can then increase the noise level inherent in the recording. Note also that variations in SNR can be

caused by head turning/movement and microphone proximity. Even in a very well recorded, highly controlled corpus such as DyViS the range of WADA SNR measurements was quite wide ranging from 10.98db for speaker 097 to 28.19db for speaker 106 (see Table 8.13).

### **8.3 Questions and Hypotheses**

Based on the research motivations and literature review, the following research questions were posed.

**Q1 To what extent does decreasing the SNR influence ASR performance on modern systems and can any tipping points be identified?**

**H1** As research demonstrated, when the SNR decreases the ASR performance will decline. A tipping point is likely when noise reaches a saturation point where it is of a similar dynamic range (or volume) to the speech. At that point it is likely that the feature extraction stage of the process will be unable to distinguish between speech and noise and the EER% will be so large as to render the system unusable.

**Q2 Are speakers with lower existing SNR/poor vocal effort affected faster, in terms of performance degradation, as the SNR incrementally decreases? Conversely, are speakers with high SNR values more resilient to the addition of noise?**

**H2** Speakers who are already exhibiting low SNR/poor vocal effort are likely to be affected to a greater extent by the incremental addition of low levels of noise than those with higher SNR.

**Q3 Does the addition of pink noise produce different results from the addition of white noise?**

**H3** Pink noise has greater energy at lower frequencies than higher frequencies (decreasing at 3dB per octave) so it should degrade speech faster than uniformly distributed white noise when added in iterative steps due to the greater quantity of voiced speech at frequencies below 4kHz which are important for ASR discrimination.

**Q4 With regard to channel matching/mismatch, is there benefit from degrading the speaker models in line with the test audio or should the speaker models be held at the highest possible quality?**

**H4** It is suggested that matching the conditions in both speaker models and test audio is likely to produce better ASR performance. However, this might not hold true for the addition of significant quantities of noise.

**Q5 With regard to the degraded results, processing plug-ins such as noise reduction and/or digital normalisation positively influence/restore ASR performance?**

**H5** The application of processing techniques, particularly noise reduction, is generally regarded as degrading the speech through spectral subtraction (i.e. removing noise will also remove some speech). It is therefore suggested that only the sparing use of adaptive NR techniques could marginally improve ASR performance for degraded data. Digital file normalisation techniques as applied using software such as iZotope might also assist with performance although it is suggested that gains are likely to be very marginal as the signal to noise ratio will remain close to the pre-normalisation levels. In other words, the overall amplitude may be boosted, but noise and signal relatively evenly so and the difference between the speech and noise (i.e. SNR) will therefore broadly remain.

## 8.4 Methodology

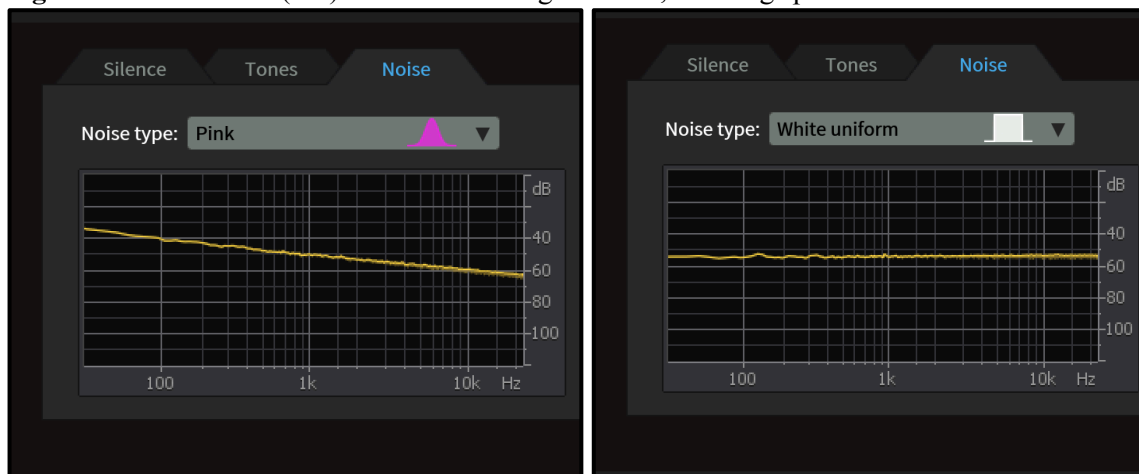
The data and equipment used was as specified in chapter 5 and Appendix G with the following adaptations.

- i. The baseline data comprised the DyViS Task 1 interview material (100 speakers SSBE, male). The net speech files were edited to generate 100 speaker models and 300 test audio files (1x minute SM and 2 x additional 1 minute TA files with the residual data comprising the third TA file).
- ii. A bespoke normative data session was created by OWR for this research experiment (i-vector/UBM, TV, LDA+PLDA session set '2016A-1024-D-CMS-Large-VAD-NoDyViS-20Apr16'). The normative data did not contain DyViS material.
- iii. The baseline and artificially degraded data was examined using OWR iVocalise version 2.1.0.1366.

It was determined that the Togneri and Pullella (2011) method of adding noise to baseline audio data was a practical and measurable way of accurately degrading the SNR under uniform and controlled conditions. Two different types of noise were selected in the signal generation plug in for iZotope RX6 Advanced [izotope.com](http://izotope.com). These were pink noise (spectrally tilting from high amplitude at low frequency to lower amplitude at high frequency) and white noise (uniform amplitude) - see Figure 8.6. These were applied to the DyViS speech baseline files using the batch facility in the iZotope application. These steps were validated with test data (i.e. addition of zero db noise) to ensure that the process itself did not influence results. The additional noise was added at 5db iterative steps beginning at -45db RMS and concluding at 0db RMS. RMS, or root mean square, refers to the averaging of the output i.e. squaring all values, determining the mean and then finding the square root of the result.



**Figure 8.6:** Pink noise (left) and white noise generators, showing spectral tilt



A total of 45 different experiments were created (Table 8.7). Each degradation step was applied to baseline files (i.e. non-cumulative). Matched/non-matched condition refers to the degradation of speaker model and test audio. Results were compared with respect to the degraded and baseline results using OWR Bio-Metrics version (v1.8.0.704) for graphing and plotting results from the .csv output files – i.e. EER% (performance), Cllr (accuracy) and LR Plots. The WADA SNR estimate for each audio file was extracted (OWR Juicer, version 2016a).

**Table 8.7:** SNR Experiments detailing noise types and settings

<b>Experiments</b>	<b>Type</b>	<b>Settings</b>
<b>Matched refers to SM&amp;TA</b>		
Baseline/control	N/A	N/A
1-10 Matched	White noise	10 iterative steps from -45 to 0 RMS
11-20 Matched	Pink noise	10 iterative steps from -45 to 0 RMS
21-30 Non-matched	White noise	10 iterative steps from -45 to 0 RMS
31-40 Non-matched	Pink noise	10 iterative steps from -45 to 0 RMS
41-43 Matched	Adaptive NR	-15db, -10db, -5db
44 Matched	Normalisation	To 0db
45 Matched	Spectral NR	-10db

For the audio enhancement experiments (41-45 inclusive) the plug-ins were selected from the iZotope RX6 Advanced suite (see iZotope.com).

Re 41-43: Adaptive noise reduction effectively learns the profile of the unwanted noise and subtracts it from the signal. It is commonly used to remove broadband or tonal noise. The settings were adjusted to ‘Advanced + Extreme’ which completes a joint time frequency analysis resulting

in fewer artefacts, but requires greater computational resources. The amount of noise reduction was varied as shown in Table 8.7. All other settings were at the manufacturers recommended positions (i.e. default).

Re 44: Normalisation (to 0db) digitally adjusts the gain across an audio file to a target peak level.

Re 45: Spectral noise reduction is similar to Adaptive noise reduction with the adaptive mode set to the off position (i.e.) constant subtraction of noise.

## **8.5 Results**

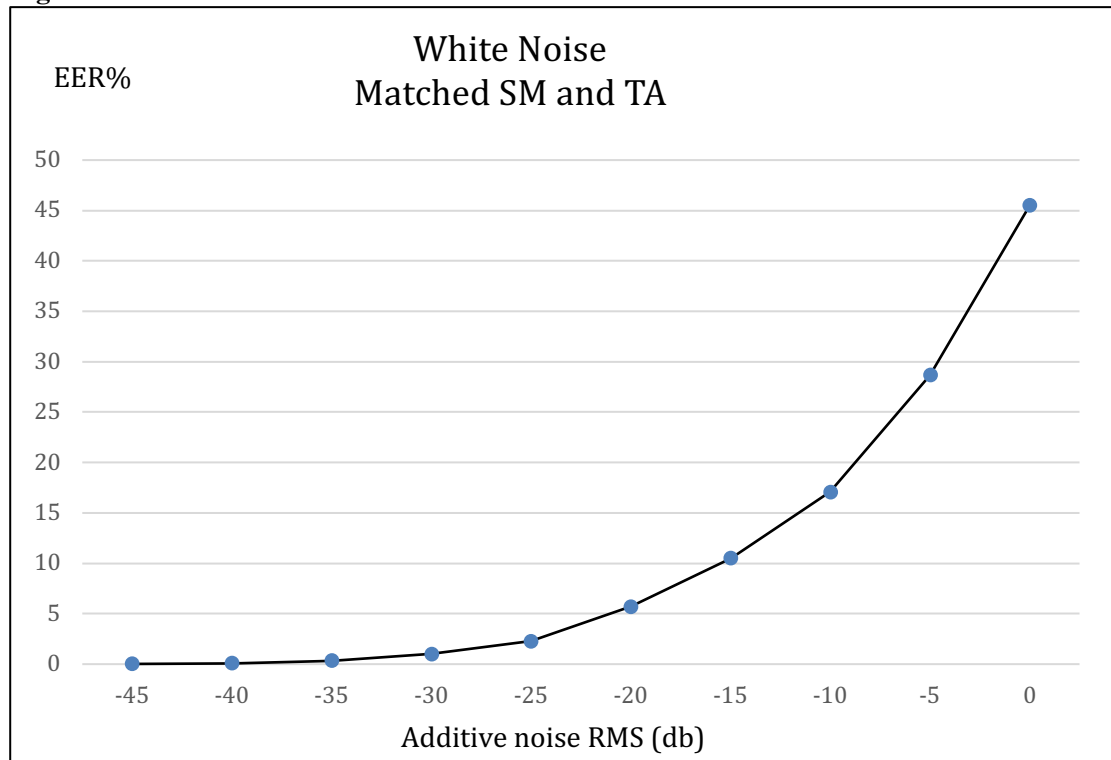
Results are presented in a series of Tables and graphs with findings discussed (8.6).

**Table 8.8:** SNR Experiments, iVocalise, i-vector/PLDA, results (next 2 pages)

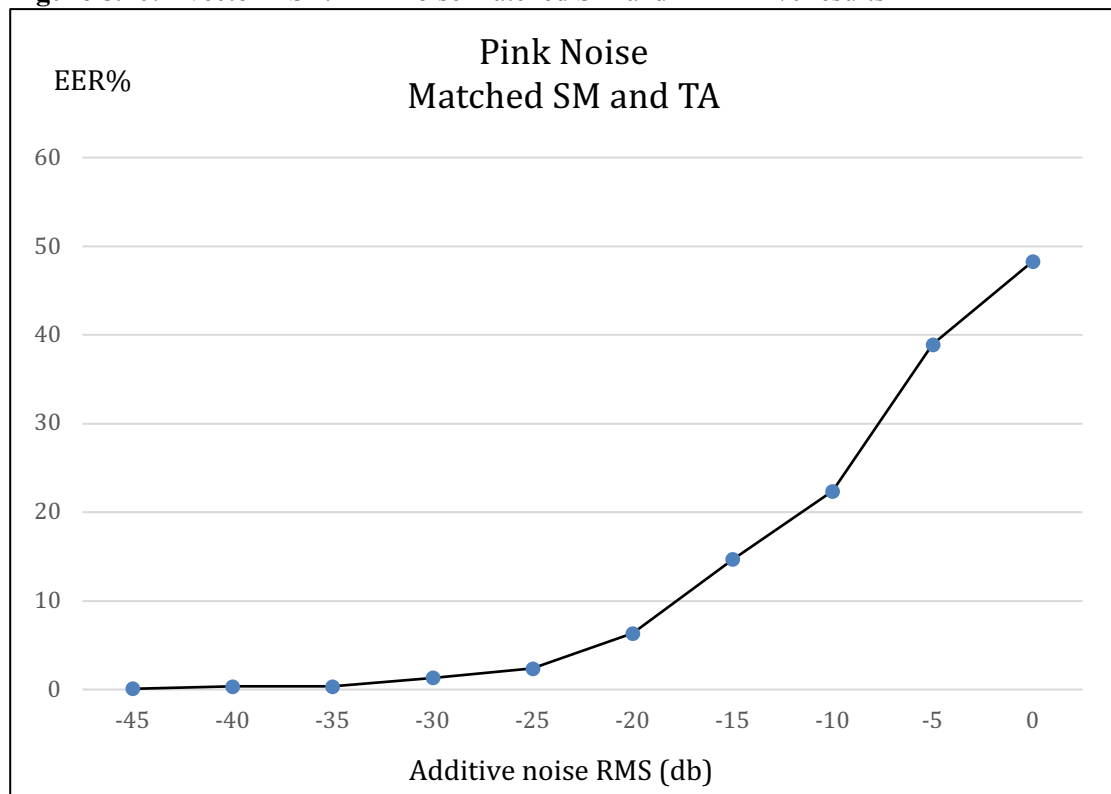
	Match SM & TA	Experiment Type [VAD off]	Noise RMS	EER	Cllr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR	FAR, FRR	FAR, FRR
										100	1,000	10,000
										%	%	%
<b>0</b>	✓	<b>Baseline</b>	<b>N/A</b>	<b>0.0051</b>	<b>0.11304</b>	<b>69.97957</b>	<b>-49.92858</b>	<b>11.94202</b>	<b>26.0617</b>	<b>0</b>	<b>0.01</b>	<b>1.33</b>
1	✓	White Noise U	-45	0.0219	0.4202	73.39019	-37.25606	11.19704	25.51934	0	0	0.33
2	✓	White Noise U	-40	0.0741	0.96509	74.29366	-28.8285	11.11599	25.4271	0	0.33	1.69
3	✓	White Noise U	-35	0.3333	2.4277	75.85648	-18.0836	10.93291	25.61115	0	1.67	22.68
4	✓	White Noise U	-30	1.0017	5.173	77.34818	-6.755244	11.38249	25.70247	0	8.48	32.35
5	✓	White Noise U	-25	2.2727	7.8867	76.76415	1.362782	12.99456	25.34248	6.67	21.43	48.67
6	✓	White Noise U	-20	5.6987	7.9518	70.59608	1.912944	16.79232	24.89449	19.33	40.03	62
7	✓	White Noise U	-15	10.4916	6.0473	57.00569	-3.495239	21.33924	25.04573	34.33	55.7	75.37
8	✓	White Noise U	-10	17.0791	6.5383	42.66234	-3.497848	22.25974	25.69852	58	86.72	95.34
9	✓	White Noise U	-5	28.6667	16.407	43.70573	18.19728	20.07989	26.88199	93.67	99.33	100
10	✓	White Noise U	0	45.4848	49.631	74.18795	68.6213	21.39767	22.76772	98	99.67	100
11	✓	Pink Noise	-45	0.0791	0.16325	71.22972	-47.90449	12.37364	26.94122	0	0.33	1
12	✓	Pink Noise	-40	0.33	0.29875	70.41334	-42.86844	12.58382	26.90666	0	0.67	2
13	✓	Pink Noise	-35	0.3519	0.66625	69.63094	-34.82043	12.6903	26.69973	0	1.67	7.01
14	✓	Pink Noise	-30	1.3333	1.3037	67.3966	-26.81032	13.44926	26.12793	1.33	4.55	17.69
15	✓	Pink Noise	-25	2.3502	1.8444	61.9313	-22.01115	15.63945	25.2174	5.67	17.43	41.01
16	✓	Pink Noise	-20	6.3283	2.4923	51.72892	-18.64746	18.87169	25.43313	23	46.07	72.35
17	✓	Pink Noise	-15	14.67	4.5616	39.49099	-11.99172	23.15419	26.35138	53.43	81.92	95.01
18	✓	Pink Noise	-10	22.3569	9.77	36.98571	3.19326	22.43772	26.89893	90.33	98.67	99.67
19	✓	Pink Noise	-5	38.9091	34.588	60.51007	46.71537	25.23389	29.43305	96.33	99.67	100
20	✓	Pink Noise	0	48.2912	66.375	93.71655	92.01427	16.79043	16.85739	97.01	99.67	100
21	×	White Noise U	-45	0.3182	0.01641	44.77537	-59.09192	13.93641	24.2753	0	1.33	2.67
22	×	White Noise U	-40	0.6818	0.21039	28.82526	-63.46144	15.81599	23.01784	0.67	4.33	12.68
23	×	White Noise U	-35	1.9731	2.313	9.250971	-68.6494	17.23731	21.67479	3.67	17.33	41.34

24	x	White Noise U	-30	4.6852	10.83	-12.29681	-74.44022	18.01885	20.38685	17.33	46.67	70.36
25	x	White Noise U	-25	10.3418	24.341	-33.34525	-80.10021	18.24501	19.27641	44.56	77	88
26	x	White Noise U	-20	17.5779	37.527	-51.97529	-85.22839	18.35684	18.41336	69.79	88.72	96
27	x	White Noise U	-15	27.5505	49.851	-69.10854	-90.73069	19.32338	17.76649	84.67	95.33	97.33
28	x	White Noise U	-10	38.2828	63.235	-87.66284	-99.98775	22.02688	18.2877	90.24	96.67	98
29	x	White Noise U	-5	46.032	77.683	-107.6915	-112.6742	22.14669	17.80822	94.67	99	99
30	x	White Noise U	0	48.5152	86.51	-119.9277	-120.6014	15.55265	14.45787	98	99.33	100
31	x	Pink Noise	-45	0.0471	0.02306	52.15806	-59.40495	13.35102	25.48679	0	0	1.33
32	x	Pink Noise	-40	0.4125	0.04315	37.56111	-64.44209	14.87607	24.42277	0	1.33	4.33
33	x	Pink Noise	-35	0.9966	0.63656	19.35217	-69.48096	16.43846	23.01376	1	5.67	17.69
34	x	Pink Noise	-30	2.6296	5.5814	-1.183407	-74.57741	17.68577	21.59637	6	29.72	46.02
35	x	Pink Noise	-25	6.633	17.834	-23.46654	-80.12043	19.35292	20.27662	31.33	56.33	76.67
36	x	Pink Noise	-20	15.7104	34.255	-47.31786	-86.40604	21.05902	18.98656	61.26	82.78	92.68
37	x	Pink Noise	-15	29.1111	50.966	-70.65434	-94.02324	22.48312	18.4728	80	91.67	96.67
38	x	Pink Noise	-10	39.096	65.906	-91.36571	-102.8277	21.61867	18.33956	93.33	97	98.67
39	x	Pink Noise	-5	45.9579	75.953	-105.2933	-109.1915	17.79203	17.33181	98.33	99.33	99.67
40	x	Pink Noise	0	49.4731	79.551	-110.2816	-110.7484	16.35793	16.19076	99	100	100
41	✓	Adaptive NR - 15db	N/A	0.0051	0.22385	74.66514	-42.46545	11.3404	25.38404	0	0	0.33
42	✓	Adaptive NR - 10db	N/A	0	0.22598	75.04484	-42.56533	11.37012	25.48847	0	0	0
43	✓	Adaptive NR -5db	N/A	0	0.20726	75.05668	-43.8732	11.43001	25.74486	0	0	0
44	✓	Normalisation to 0db	N/A	0.0017	0.15105	74.37363	-48.01534	11.68433	26.35088	0	0	0
45	✓	Spectral Denoise - 10db	N/A	0.0135	0.35387	75.46694	-37.20133	11.35659	24.63742	0	0	0.67

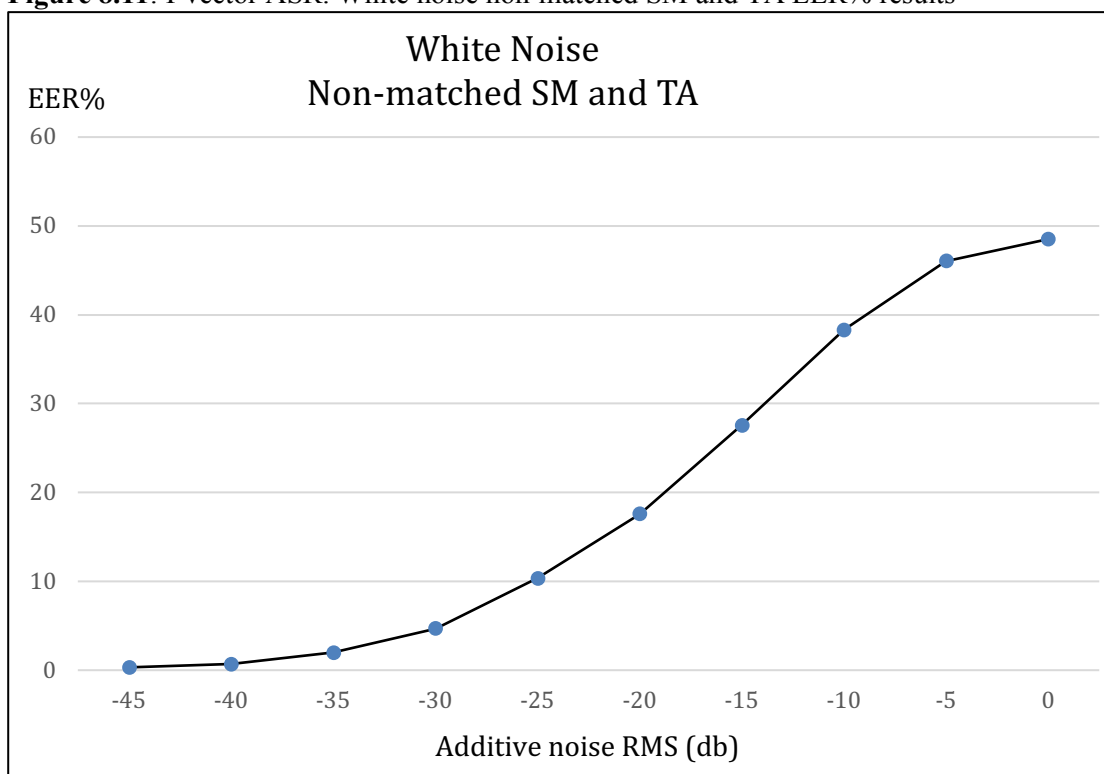
**Figure 8.9:** I-vector ASR. White noise matched SM and TA EER% results



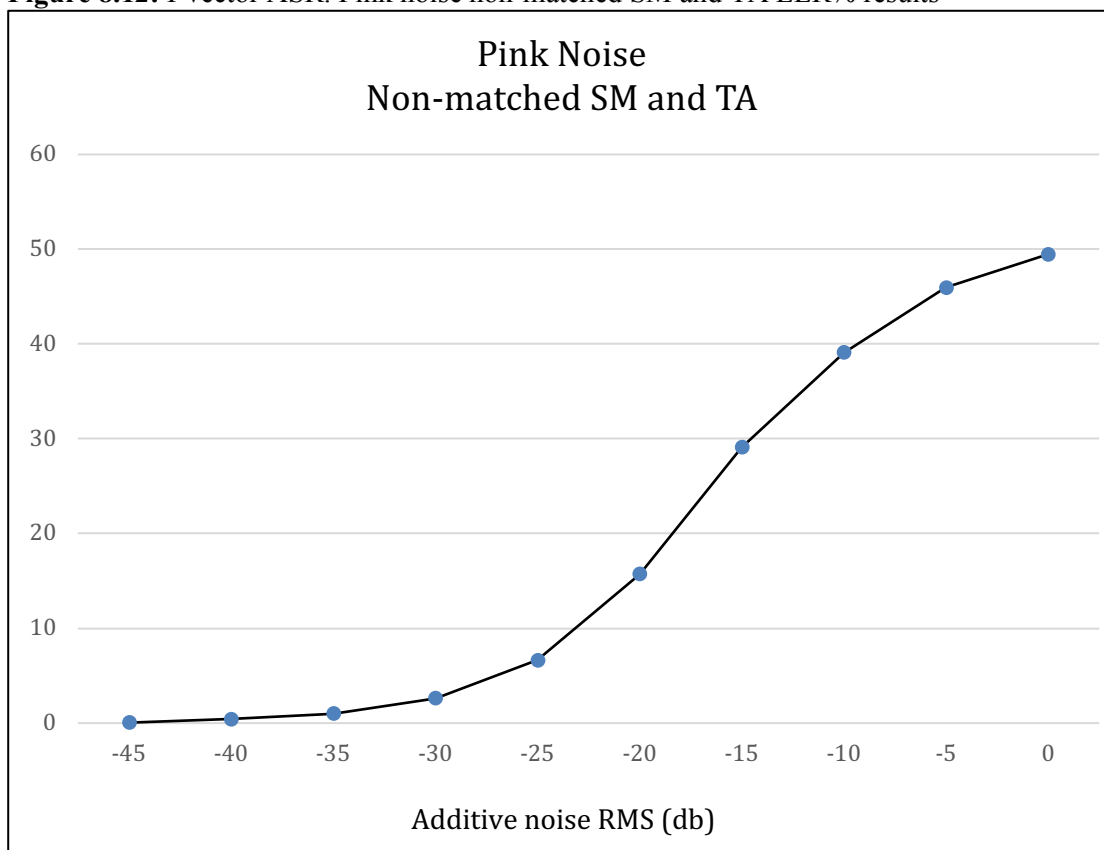
**Figure 8.10:** I-vector ASR. Pink noise matched SM and TA EER% results



**Figure 8.11:** I-vector ASR. White noise non-matched SM and TA EER% results



**Figure 8.12:** I-vector ASR. Pink noise non-matched SM and TA EER% results



**Table 8.13:** WADA SNR Estimates for 100 x DyViS speakers, task 1 (SM, baseline)

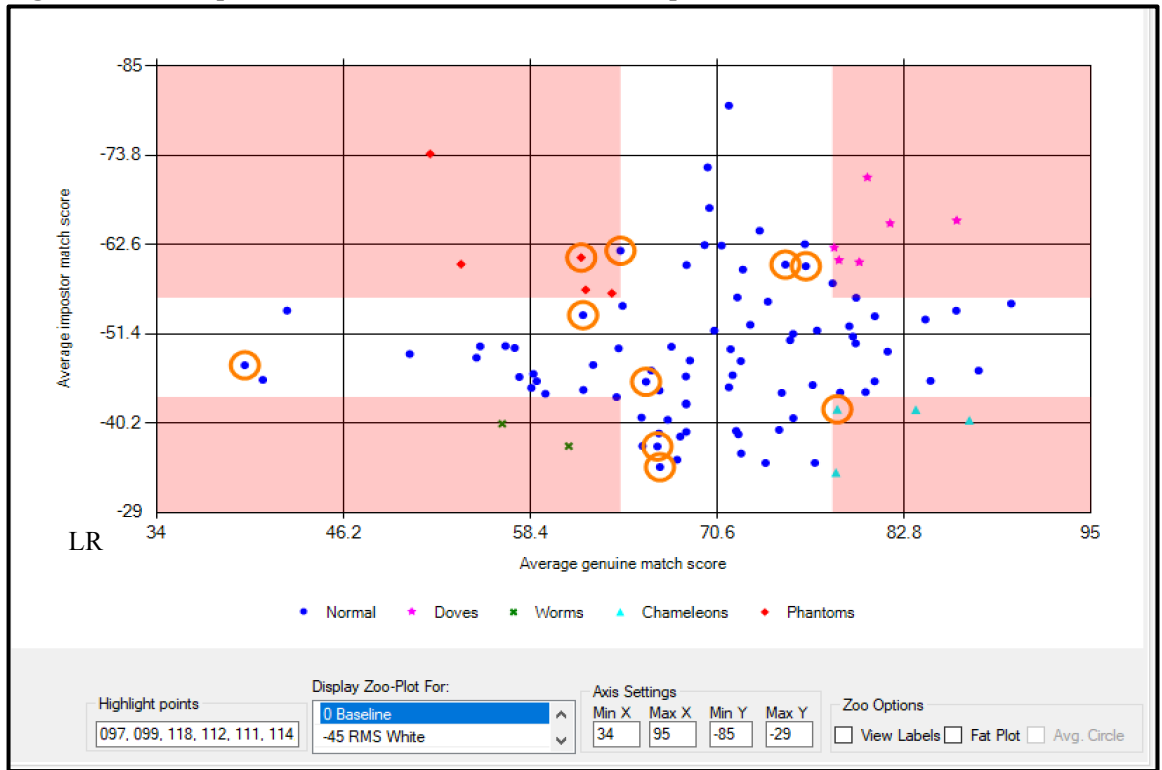
DyViS Speaker	WADA SNR (db)	DyViS Speaker	WADA SNR (db)	DyViS Speaker	WADA SNR (db)	DyViS Speaker	WADA SNR (db)
001	20.209	029	20.588	056	16.620	086	19.611
002	19.532	030	24.545	058	21.793	087	19.656
003	16.764	031	18.777	059	26.447	088	20.232
004	16.371	032	23.196	060	17.843	090	19.964
006	19.504	033	25.580	062	24.322	093	15.384
008	19.689	034	19.032	063	22.626	094	20.783
009	20.237	035	14.394	064	21.502	095	20.056
010	18.056	036	19.575	065	16.242	096	18.300
011	20.677	037	16.792	066	21.253	097	10.978
012	21.551	038	18.159	067	14.611	099	11.825
013	16.675	039	22.942	068	20.829	100	18.357
015	22.397	040	15.247	069	15.868	102	21.575
016	19.119	042	13.802	071	22.864	103	18.008
017	17.428	043	20.825	072	19.764	105	15.977
018	22.577	044	20.028	073	18.903	106	28.190
019	24.581	045	20.811	074	16.170	107	17.280
020	16.533	046	16.358	075	15.810	108	16.720
021	21.761	047	21.261	076	23.162	111	12.852
022	20.104	048	21.633	077	26.167	112	12.842
023	27.001	049	21.453	078	20.337	113	18.162
024	16.658	050	16.307	079	19.776	114	13.010
025	18.476	051	18.780	080	20.693	115	15.127
026	22.736	052	17.580	081	15.475	118	12.822
027	18.467	053	20.383	084	22.488	120	14.882
028	19.367	054	18.231	085	17.902	121	14.643

Top 10% WADA SNR (highest db first) 106, 023, 059, 077, 033, 019, 030, 062, 032, 076

Bottom 10% WADA SNR (lowest db first) 097, 099, 118, 112, 111, 114, 042, 035, 067, 121.

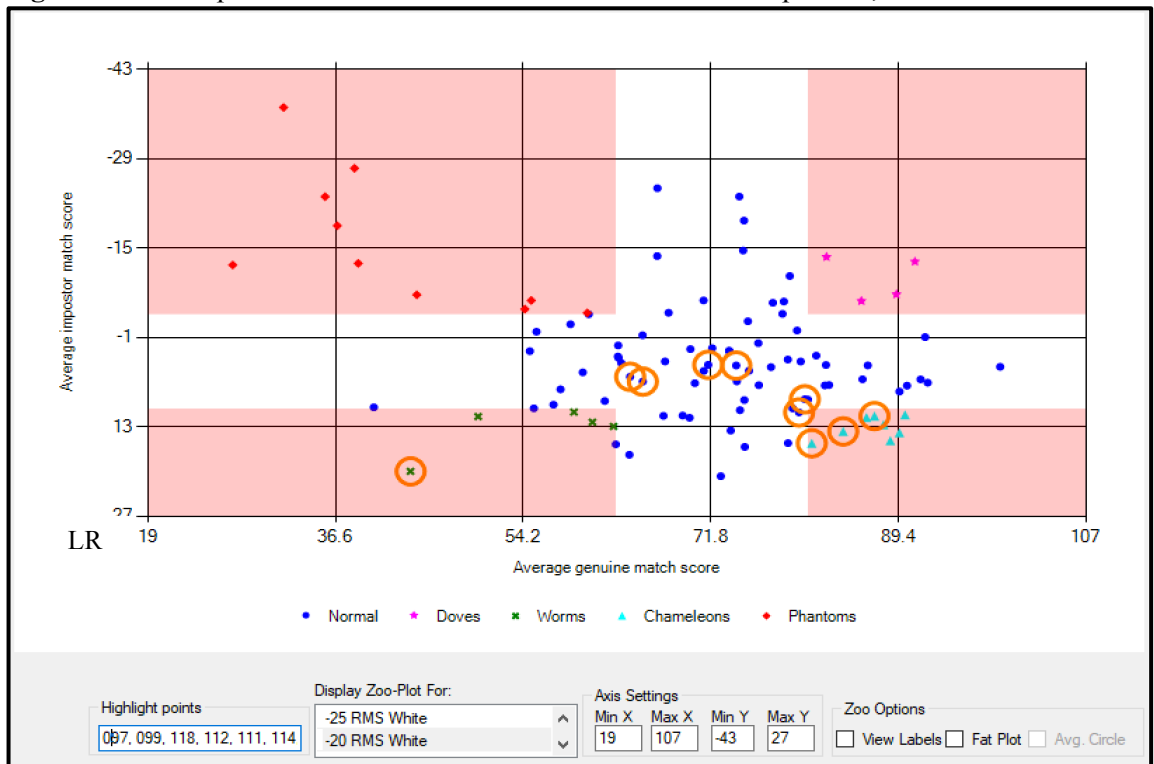
The speakers above are circled in the zoo plots below (Figures 8.14, 15, 16 and 17). All other speaker labelling has been deliberately removed to enable improved viewing across all results.

**Figure 8.14:** Zoo plot baseline results. Lowest 10% of speakers, WADA SNR



Speakers with the lowest WADA SNR ratings (Figure 8.14) were not grouped or positioned in the poorer performing quartiles (worms, phantoms and chameleons) in the baseline results, but distributed broadly in the central range. With the addition of noise (Figure 8.15), speaker positions appeared to cluster towards the lower right (chameleons) indicating performance degradation.

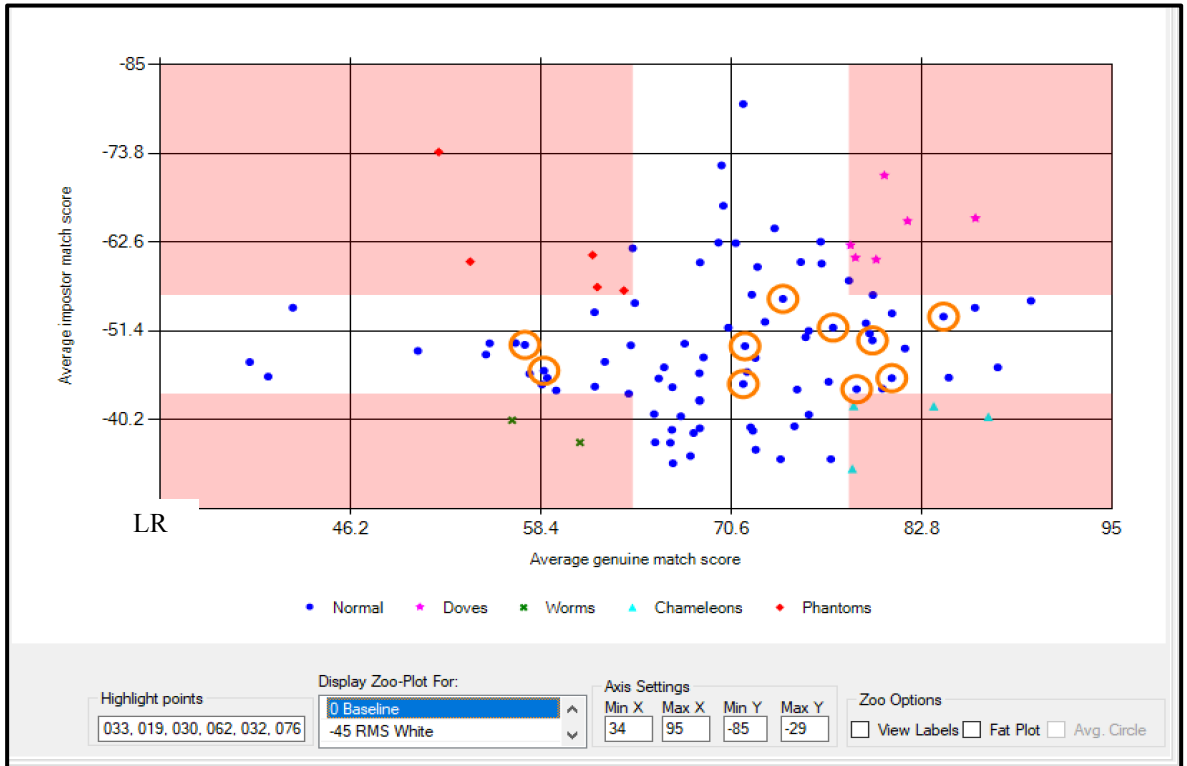
**Figure 8.15:** Zoo plot -20db RMS White Noise. Lowest 10% of speakers, WADA SNR





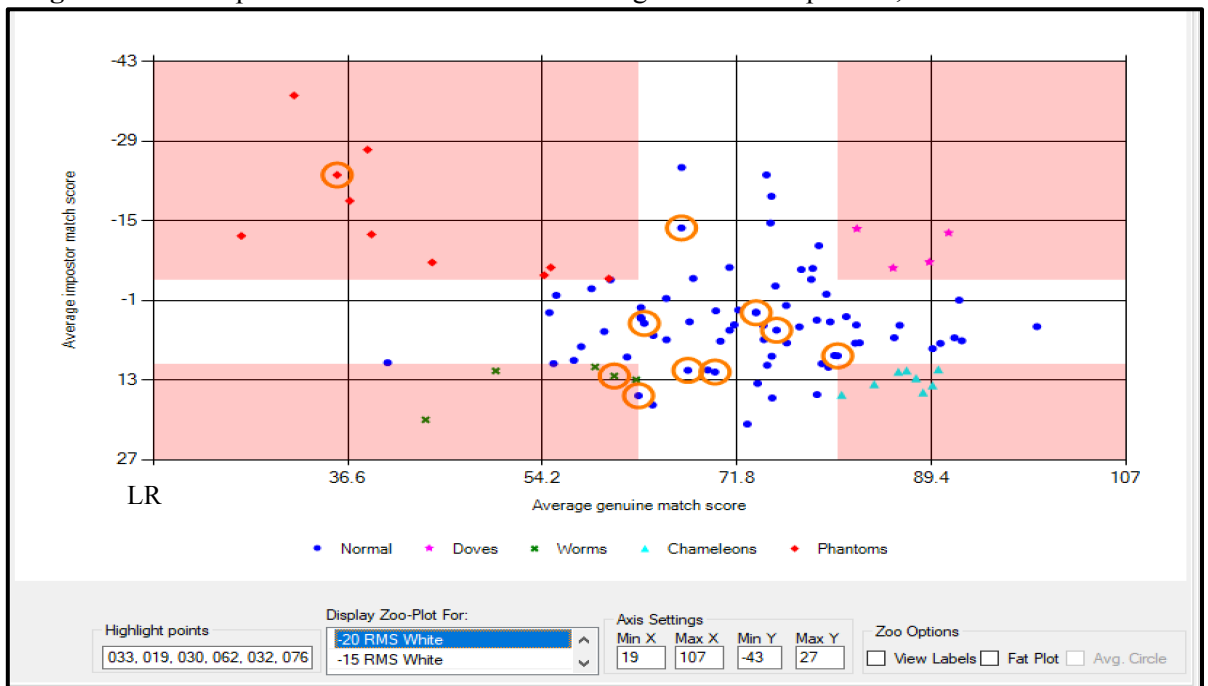
At an approximate tipping point of -20db RMS, higher imposter scores and higher genuine scores were noted - with clustering towards the Chameleon quadrant. Note also the difference in axis numbering to Figure 8.14, indicating overall scores, and a higher quantity of phantoms (red).

**Figure 8.16:** Zoo plot baseline results. Highest 10% of speakers, WADA SNR



Speakers with the highest WADA SNR ratings (Figure 8.16) were not all positioned in the Dove quartile in the baseline results, but distributed broadly in the central range with potential clustering towards the right (marginally higher genuine match scores).

**Figure 8.17:** Zoo plot -20db RMS White Noise. Highest 10% of speakers, WADA SNR



With the addition of noise (Figure 8.17) speaker positions moved towards the lower left indicating performance was negatively influenced. Note also the difference in axis numbering to Figure 8.16 – demonstrating overall score movement. At an approximate tipping point of -20db RMS, higher imposter scores and higher genuine scores were noted, but the movement towards the Chameleon quadrant was not as noticeable as in Figures 8.14 and 8.15. Of the overall chameleons in the degraded data, 50% of them were in the lowest 10% WADA SNR. This suggests the difference in speaker performance is marginally dependent on the initial WADA SNR values (i.e. those speakers with lower initial SNR are marginally more prone to the addition of noise).

## 8.6 Findings

In response to the research questions (8.3).

**Q1 Recap**      **To what extent does decreasing the SNR influence ASR performance on modern systems and can any tipping points be identified?**

**A1**              As predicted and in line with previous research ASR performance declined as noise was added and the SNR decreased and proportionate to the quantity of noise added. For non-matched conditions (SM and TA) performance was affected to a greater extent. The EER% practically doubled for every 5db increment in noise (or 5db decrease in SNR). For the addition of lower levels of noise (-45 to -30db) the effect on EER% was therefore noticeable but relatively small due to the doubling of a very low number. However, performance rapidly decreased at -25db to -15db (approximate tipping point). With only one exception, for which EER% was extremely high, Cllr (accuracy) declined on the addition of any noise. As seen in Table 8.8, mean H0 values decline and H1 values rise in comparison to baseline, causing the overlap between H0 and H1 distributions to increase. As predicted very high levels of noise, where the speech was barely audible, rendered the system unusable (EER close to 50%).

**Q2 Recap**      **Are speakers with lower existing SNR/poor vocal effort affected faster, in terms of performance degradation, as the SNR incrementally decreases? Conversely, are speakers with high SNR values more resilient to the addition of noise?**

**A2**              Speakers with initially low WADA SNR/lower vocal effort were not noticeably affected faster by the incremental addition of noise than those with higher baseline WADA SNR. Zoo plots highlighted the top 10% and lowest 10% speakers (Figure 8.14 to 8.17 inclusive) and showed that there were no strong correlations with speaker performance position. A plausible explanation for this could be that adjustment for this is made, to some extent, in the pre-emphasis phase of the MFCC extraction (3.4.3). Nevertheless, speakers with lower initial WADA SNR did have a marginal tendency to move towards the lower right quartile (larger number Chameleons) as SNR decreased, than those with the higher initial WADA SNR. The results are far from conclusive

since baseline zoo plots for highest and lowest 10% did not provide definitive separation in terms of clustering/positioning.

**Q3 Recap** Does the addition of pink noise produce different results from the addition of white noise?

**A3** As posited, pink noise degraded ASR performance faster than uniformly distributed white noise when added in iterative steps. As suggested, this is likely due to the greater influence on the lower frequencies of speech.

**Q4 Recap** With regard to channel matching/mismatch, is there benefit from degrading the speaker models in line with the test audio or should the speaker models be held at the highest possible quality?

**A4** Matching the speaker models to the test audio files provided the best EER% in all instances.

**Q5 Recap** With regard to the degraded results, can processing plug-ins such as noise reduction and/or digital normalisation positively influence/restore ASR performance?

**A5** As hypothesised the sparing application of (iZotope RXAdvanced) adaptive noise reduction marginally improved ASR performance under matched conditions (Table 8.18) with performance increases noted (e.g. from 0.0051 to 0.0000 EER%). Digital normalisation also assisted performance to a very small extent too (0.0051 to 0.0017 EER%). The application of a much higher quality adaptive spectral noise reduction (-10db) plug in, however, provided a performance decrease (0.0051 to 0.135 EER%). Cllr (accuracy) rose, very marginally, in all instances.

**Table 8.18:** Summary of results from audio enhancement experiments

Treatment	EER%	Cllr
Baseline	0.0051	0.11304
Adaptive NR -15db	<b>0.0051</b>	<b>0.22385</b>
Adaptive NR -10db	<b>0</b>	<b>0.22598</b>
Adaptive NR -5db	<b>0</b>	<b>0.20726</b>
Digital Normalisation to 0db	<b>0.0017</b>	<b>0.15105</b>
Spectral Denoise -10db	<b>0.0135</b>	<b>0.35387</b>

**Red** = Poorer, compared to baseline

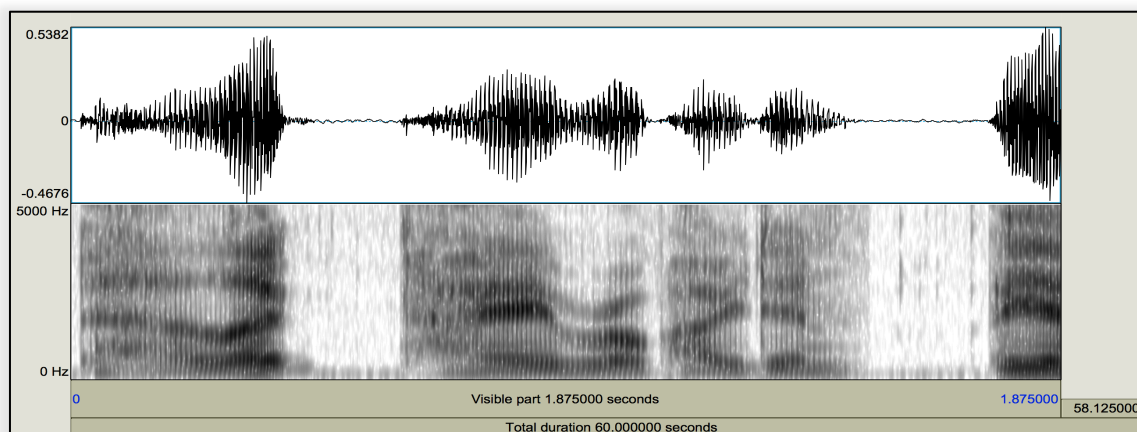
**Green** = Improvement, compared to baseline

Further research is required to establish the application of processing techniques as audio enhancement can clearly produce unpredictable results and increasing EER% performance whilst decreasing accuracy would also not be recommended.

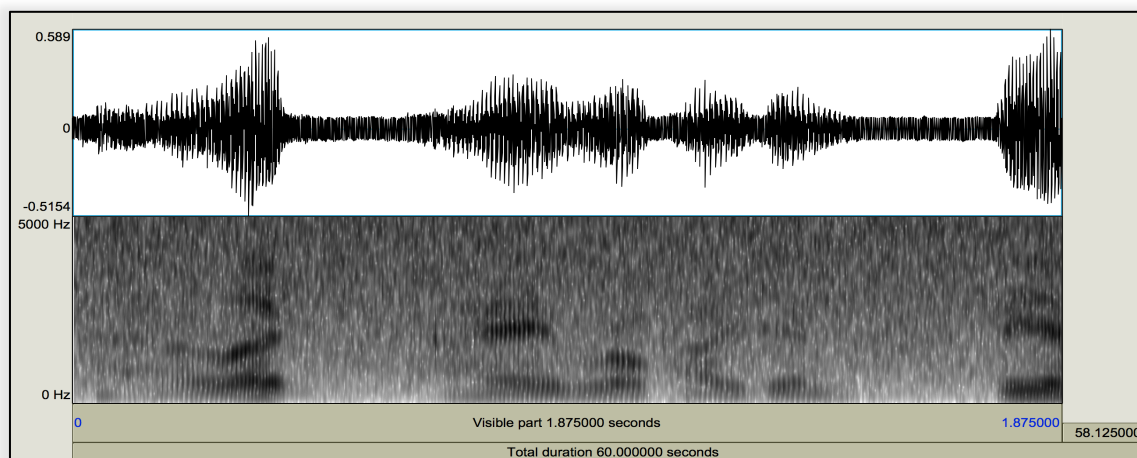
## 8.7 Discussion

Despite improvements to the underlying architecture of ASR systems (i-vectors) results broadly supported previous research. The experiments demonstrated that SNR continues to have a significant influence on ASR performance, despite advances from GMM-UBM to i-vector approaches. Thresholds and tipping points were determined. Whilst it is conceded that these are specific to the i-vector ASR system used, normative data (UBM, TV, LDA+PLDA) and the SNR inherent in the baseline data is relatively high, it is hoped that the Tables provided assist in informing casework analysis and draw attention to the issues of poor SNR/vocal effort speaker comparison. Note the addition of noise in Figure 8.20 in comparison to Figure 8.19 (at the default spectrogram dynamic range of 70db). A plausible explanation for the degradation in performance is that noise simply interferes with the feature extraction process and, therefore, in the statistical modelling phase.

**Figure 8.19:** Praat spectrogram. Speaker 2 SM (1.875s) baseline



**Figure 8.20:** Praat spectrogram. Speaker 2 SM (1.875s) -20dbRMS white noise



When examining the degraded audio files it was understandably difficult to hear speech content over noise greater than -25db RMS. To some extent it was then remarkable that any meaningful ASR results were produced beyond -10db RMS at all. Speech became practically inaudible and,

from purely a subjective perspective, the threshold for both acoustic and auditory analysis appears similar in scale - strongly supporting the use of quantitative technical assessment prior to ASR examination.

Finally, future research to assess whether a forensic examiner's EER% (i.e. auditory analysis) on degraded speech would generally compare more favourably against that of an ASR. Whilst difficult to measure, and apply across casework analysts, research could assist in determining when an auditory approach should be completed in preference to ASR analysis (or vice-versa).

# Chapter 9 Reverberation

---

This chapter examines the influence of reverberation on ASR system performance. The main objectives were to examine the extent of performance degradation and to inform the application of ASRs in comparison casework. The chapter begins by discussing the research context and provides a literature review of the publications that informed methodology and the subsequent experiments conducted. An overview of the research methodology is then given and an outline of the experiments conducted is provided.

The research questions are then stated, with associated hypotheses. Ten reverberant conditions are modelled in software and then applied to the baseline data. Both matched (equivalently degraded) and unmatched (non-equivalently degraded) conditions are tested in the context of speaker models and test audio files.

The data is then passed to two different ASR systems for comparison, a GMM-UBM OWR Vocalise system (2013) and a later i-vector (UBM, TV, LDA+PLDA) OWR iVocalise system (2017). A bespoke UBM is created specifically for the experiments and a specially adapted iVocalise/PLDA (session 1) is utilised to ensure normative data relevance and optimise performance. Variations are then also made in the normative data, for the i-vector/ system, with two additional bespoke UBM, TV, LDA+PLDA (sessions 2 and 3). The experiments are re-run to provide further ASR performance comparisons between GMM-UBM and i-vector systems against the baseline results and between the three different UBM, TV, LDA+PLDA sessions. Detailed analysis is then provided using biometric graphs to illustrate the influence of reverberation on ASR performance.

The chapter concludes with discussion placing the results in the context of the initial research conducted and then in the wider field of FSC. Several practical recommendations are made concerning handling field recordings that contain reverberant speech and future research recommendations are made.

## 9.1 Introduction

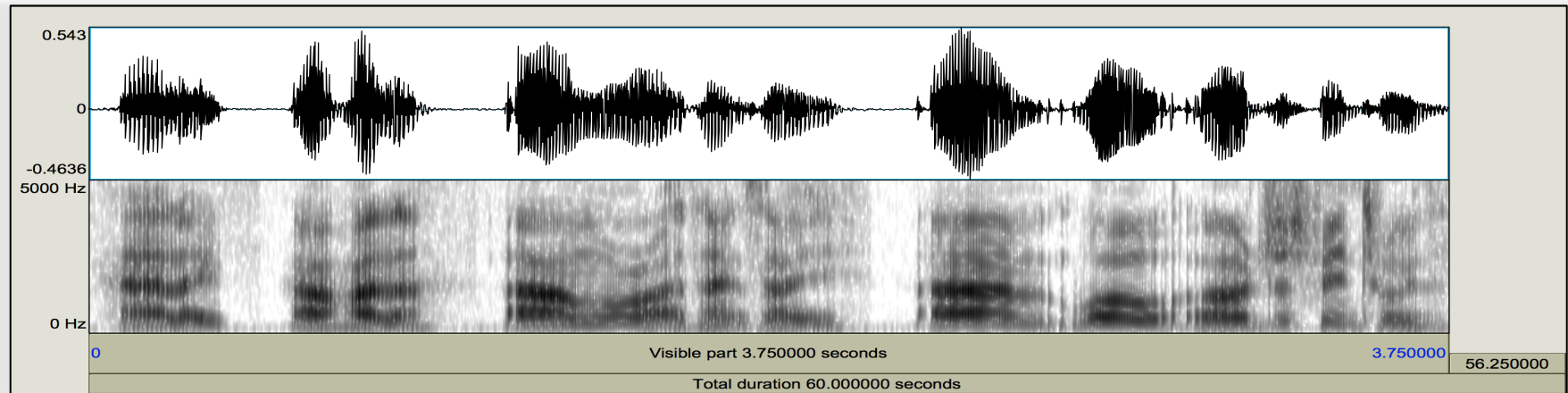
Yoshioka et al. (2012) defined reverberation as ‘the repeated sound reflections in a room (which) create a sequence of numerous slowly decaying copies of the original sound.’ (2012: p.116). Whilst their paper focused primarily on the challenges of reverberant speech for content recognition purposes (i.e. speech to text and machine intelligibility) rather than speaker comparison, the underlying principles in room acoustics and challenges with regard to reverberation noise are very

similar, since almost identical feature extraction methods are employed. Their study and the speech recognition Reverb Challenge (2014) assisted in guiding the research experiments in this thesis in addition to those relating specifically to speaker verification system reverberation research.

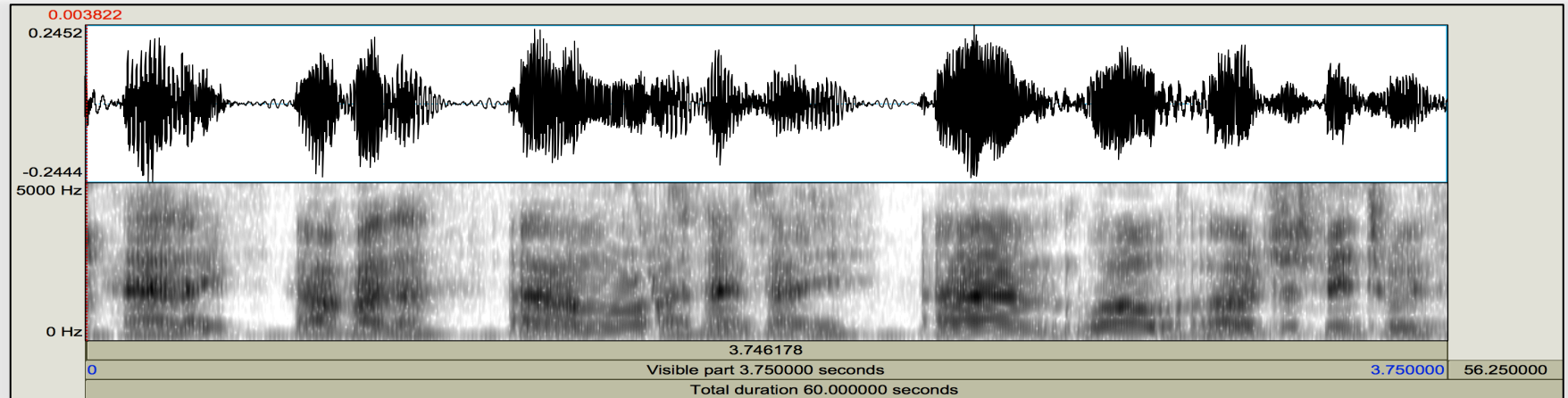
An important and common measurement of reverberation is referred to as RT60. This is effectively the length of time that it takes for a reverberant sound to reduce by 60db. The idea of measuring impulse and response characteristics was pioneered by Schroeder (1964), who used filtered pistol shots to measure the response times in reverberant rooms. As described in 3.3.7, direct sound arrives slightly before the first reflections (early sound). The delay between the two is referred to as ‘pre-delay’ or the ‘initial time gap’ (Dario and Barbosa, 2012).

In reverberant conditions, early (approximately 50-80ms) and late reflections (>80ms) merge with the direct speech signal causing a self and overlapping masking effect (Sadjadi and Hansen, 2010). In essence, it is these reflections combining with the direct sound that causes the speech to ‘smear’ in the time domain. This is both audible, and visible in a spectrogram as presented in Yoshioka et al. (2012) and Sadjadi and Hansen (2010). Four spectrograms, taken from the research conducted in this chapter, are presented by way of example. Figures 9.1 to 9.4 (next 2 pages) show the spectrogram and waveform views of a typical DyViS speaker model utterance. Note that all visualisations display the same edited extract taken from the first 3.75s of speech from DyViS speaker number 120 (edited).

**Figure 9.1:** Baseline data, no reverberation applied. DyViS Speaker 120

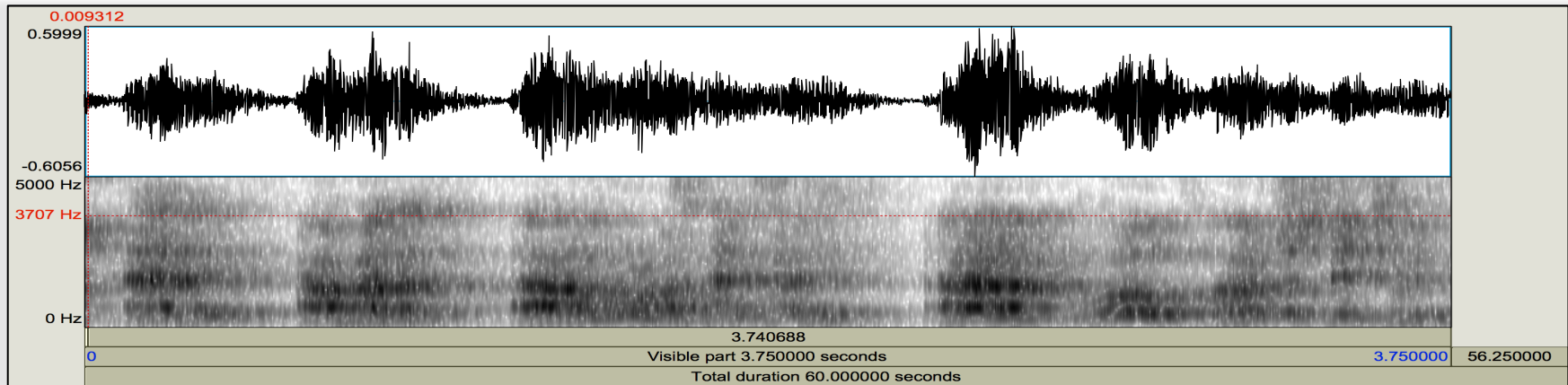


**Figure 9.2:** Car reverberation applied (RT60 = 0.60). DyViS Speaker 120

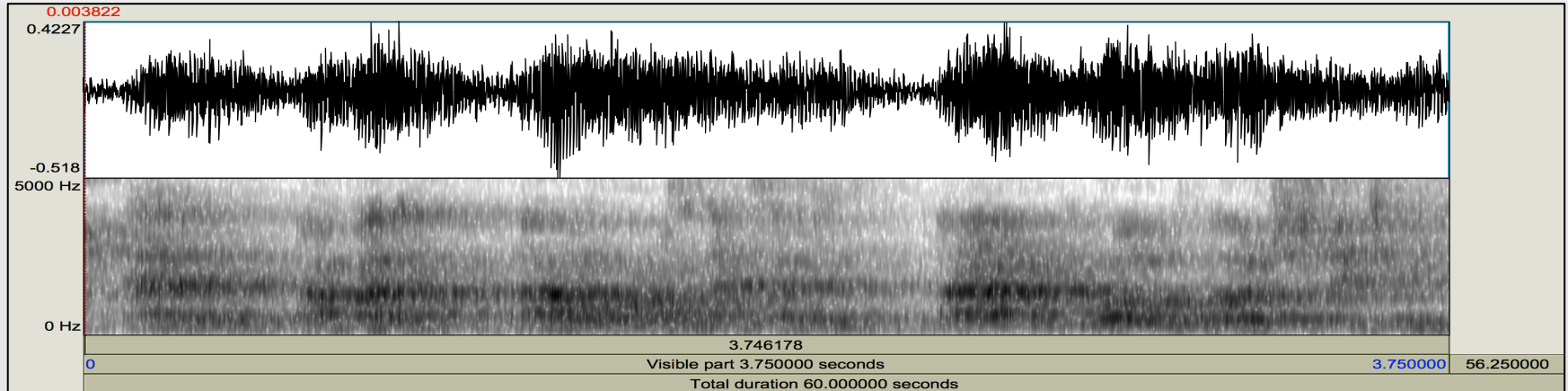




**Figure 9.3:** Living Room reverberation applied (RT60 = 0.70). DyViS Speaker 120



**Figure 9.4:** Hall reverberation applied (RT60 = 1.40). DyViS Speaker 120



The comparative images presented (Figures 9.1 to 9.4) show the degradation, or spectral smearing, caused by reverberation with the following observations:

- i. As the size of the room, and RT60, increases the speech degradation becomes more severe. Phoneme boundaries become progressively blurred. Small pauses fill with reverberant noise
- ii. The glottal pulses, clearly visible in the vertical plane of the spectrogram (untreated baseline data, Figure 9.1) – these quickly lose definition in subsequent spectrograms.
- iii. Bilabial plosives (/p, b/), velar plosives (/k, g/) and fricatives (/s, ʃ, θ/) that are usually represented by relatively short, bursts of energy in the vertical plane of the spectrogram (e.g. Figure 9.1). These become dispersed (in time) as RT60 rises, noticeably smeared to the right as RT60 further increases and then are effectively lost as they effectively merge with adjacent phonemes (e.g. Figure 9.3, RT60 = 0.70).

Whilst sound travels at a relatively constant speed of approximately 343m per second, dependent on air temperature and humidity, reflections from close proximity and distant surfaces arrive at the listener or microphone at different times (see 3.3.7 re RT60, direct and early sound). Further complex sound reflections arise as surfaces have different absorption/reflection properties. This may involve environmental variability caused by building materials, wall surfaces, ceiling and floor coverings, windows, soft and hard furnishings and even any people present. The complex influence of reverberation degradation on ASR performance is therefore the focus of this chapter.

## 9.2 Literature Review

The study of ASR system performance under reverberant conditions is well researched. This section places the present research conducted in this thesis in the wider context of previous research. The influence of reverberation on the intelligibility of speech is also well researched, but considered out of scope for this thesis.

Castellano et al. (1996) demonstrated that ASR performance degraded sharply for reverberant speech. Their research sought to mitigate this through the treatment of training material (speaker models) with similar reverberation to the test material. Attempting to match conditions in this way demonstrated a 5.45% performance decrease against baseline data, compared to a 13.7% decrease for unmatched conditions.

Zhao, Wang and Wang (2014) examined the combined issues of environmental noise (referred to as ‘factory’ and ‘engine room’) and reverberation. Their study took 300 random speakers from the NIST 2008 SRE data, degraded the audio under controlled conditions and then established ASR performance (GMM-UBM) against a baseline system. The group developed a two-step approach

to problem solving. The first stage attempted to remove the background noise using a deep neural network (DNN) classifier and binary masking. The second stage reduced the influence of degradation through 'deliberately introducing reverberant noise to speaker models in order to reduce the mismatch' (2014: p.836). This they completed by capturing real world impulse responses from four microphone positions to induce T60 values, rather than using DSP/plugin-ins. The group presented results in SID accuracy % and it was found that even when degrading through reverberation alone performance reduced from 97.3% (anechoic set) to 77.08% (reverberant set). In attempting to then match reverberant conditions (speaker model and test audio), the group were able to restore accuracy somewhat, achieving an optimum of 86.00% for 600ms (RT60). They concluded their findings by stating that training speaker models in multiple conditions of reverberation could improve ASR performance. However, it could be argued that this might be overestimating the simplicity of reverberation, which can include complex reflections and frequency dampening.

The Zhao group also conceded that the feature extraction process itself became more problematic when dealing with reverberant speech. Feature extraction pertaining to reverberant speech was further examined by Ganapathy, Pelecanos and Omar (2011) - in the similar context of speech recognition. The first group demonstrated that, by extending MFCC windowing values beyond the values of RT60, they were able to improve performance by relative values of 20-30%. Mitra, Franco and Graciarena (2013) exchanged the feature extraction method altogether - demonstrating a broad, if small, performance increase by using DOCC (dampened oscillator cepstral coefficients) in preference to MFCCs. Shabtai, Rafaely and Zigel (2010) further examined the feature extraction process and the effect of reverberation on GMM ASRs. Their research took 14 different environments and applied reverberation to establish the influence of temporal smearing. They concluded that performance decreased as RT60 increased. For audio that had particularly high RT60 values the group also recommended that cepstral mean subtraction (CMS) (Furui, 1981; 2001) was not used by default as they found that, in some instances, CMS caused EER% to rise. These works further influenced the forming of research question 3b in this thesis - examining improvements that could be made to the feature extraction process and ASR system/UBM, TV, LDA+PLDA configuration.

Peer, Rafaely and Zigel (2008) also demonstrated significant ASR performance degradation caused by reverberation. They suggested several ways of mitigating the effect including score normalisation, adjusting the background model and attempting to match the acoustic conditions by training bespoke models. Akula and De Leon's study (2008) and a study by Akula, Apsingekar and De Leon (2009) also found that reverberation degraded ASR performance by up to 30%. However, by capturing the IR of the original recording environment, in which the suspect recordings (TA) were made, they were able to effectively remodel and then compensate for

conditions by applying the IR to the speaker models. This then invoked a (replicated) channel match between SM and TA. Their method successfully achieved 30%, 22% (small offices), 16% (lounge), 13% (conference room) relative improvement in performance (clearly dependent on room size). Zieger and Omologo (2008) also successfully demonstrated that impulse responses applied to clean speaker models can improve performance on contaminated test audio. They applied a fused model methodology (clean and contaminated) to 40 speakers and demonstrated slightly smaller performance gains at 0.61% (average EER decrease on test audio).

Applying reverberation to improve channel matching was a common theme in research studies and informed the methodology of experiments in this chapter. However, the problem of convolution noise is extremely complex and clearly cannot be completely mitigated by a single adjustment. There are also significant practical challenges in completing an artificial channel matching process for forensic speaker comparison casework:

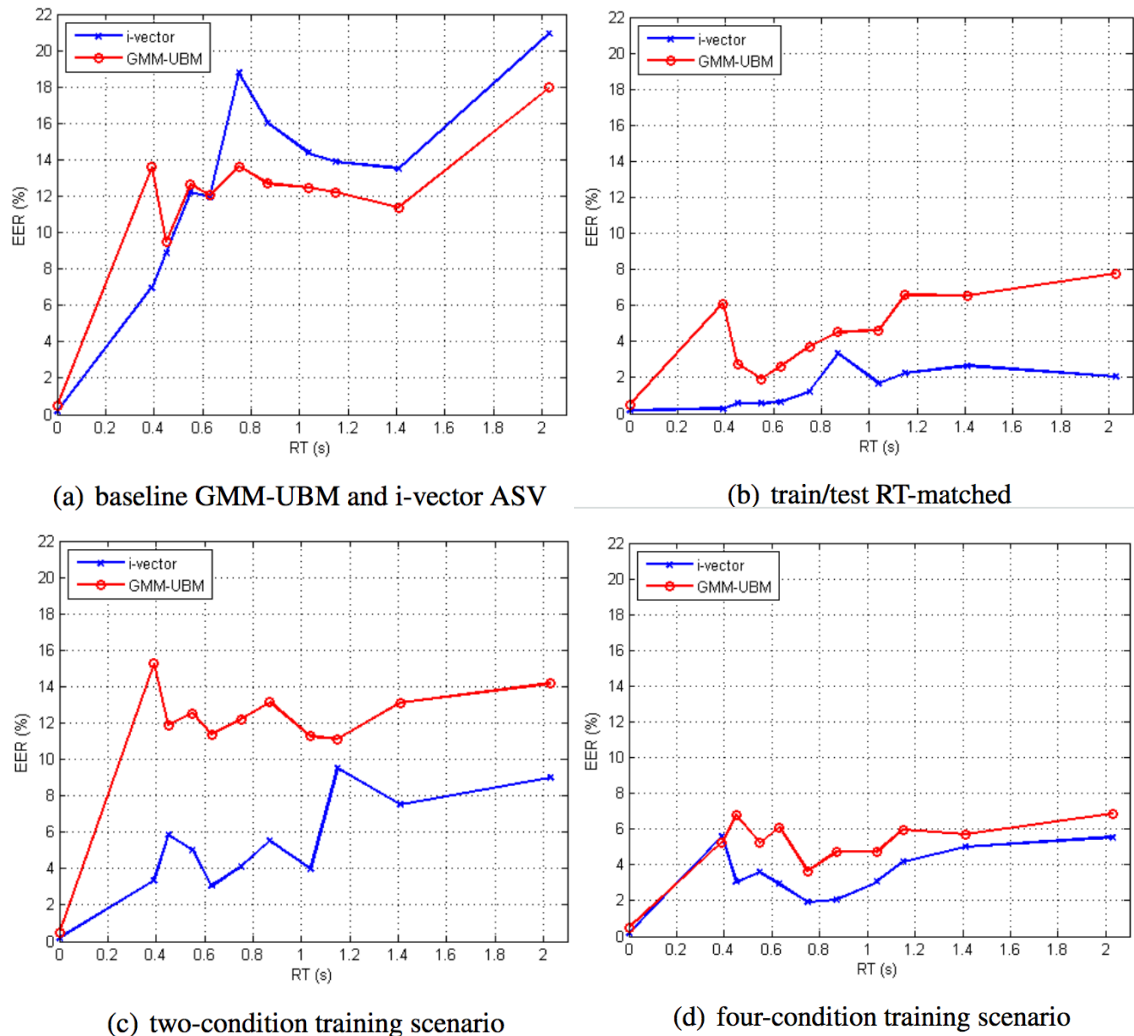
- i. Capturing an impulse response measurement from the exact location of the suspect recording in the same space (i.e. proximity, room/furniture).
- ii. Accurately measuring the complex reverberation settings inherent in the unknown speaker recording and then correctly applying the same settings to the speaker model(s) artificially is likely to introduce unknown variables and a system that cannot be validated.
- iii. Amendment of the normative data of the ASR (UBM, TV, PLDA) is likely required, but it is difficult to know as to what to amend it to.
- iv. Quantifying and recalibrating a 'new' system (score height, LLR thresholds, EER%, accuracy, precision) and then demonstrating best practice for forensic standards.

Ming, Hazen, Glass and Reynolds (2007) and Garcia-Romero et al. (2012) drew attention to the performance difference between clean speech analysed in ASR systems, as opposed to speech from 'real world' conditions. The latter demonstrated that constructing more complex normative/PLDA data (see chapter 3) from capturing several multiple reverberant conditions, could improve ASR performance in some instances.

Avila et al. (2015) were the first to publish comparisons between ASR GMM-UBM and i-vector systems (UBM, TV, LDA+PLDA) in relation to reverberation. Their study researched performance variation under four different training conditions (matched and unmatched) across the two types of Microsoft Speaker Recognition (MSR) systems when reverberant noise was added (36 speakers, read speech). Interestingly, the group down-sampled their speech material to 8kHz sample rate before completing the feature extraction stage – which could have negatively influenced EER%, see results in chapter 10. Full details of the construction of UBM and UBM, TV, PLDA configurations were not provided. Results from the Avila group study found that performance degraded for reverberant conditions. They also found that matched RT60 conditions performed

optimally, for both engines and that the i-vector system generally outperformed the GMM system under reverberant conditions with a small anomaly for baseline conditions. However, the difference in performance between the two systems was relatively small, particularly for the 4-condition experiments. Results obtained (Figure 9.5) showed close performance alignment for both systems at  $RT60=0.4s$  and  $RT60=0.6s$  (approximately 6% and 7% EER respectively).

**Figure 9.5:** Summary of results reproduced from Avila et al. (2015: p.4)



The Avila group recommended multi condition training data based on their 4-condition configuration results and improving the quantity of reverberant data in the training (and normative set) is a logical progression. Nevertheless, the EER% performance is never likely to be as good as baseline and  $RT60$  is a rarely quantified variable in field comparison and therefore almost impossible to practically or accurately reproduce.

Shabtai, Rafaely and Zigel (2010) suggested that when  $RT60$  is greater than the short-term Fourier transform frame size then time smearing will occur in the extracted feature vectors. As  $RT60$  increases then this effect worsens and the means calculated for the statistical modelling phase

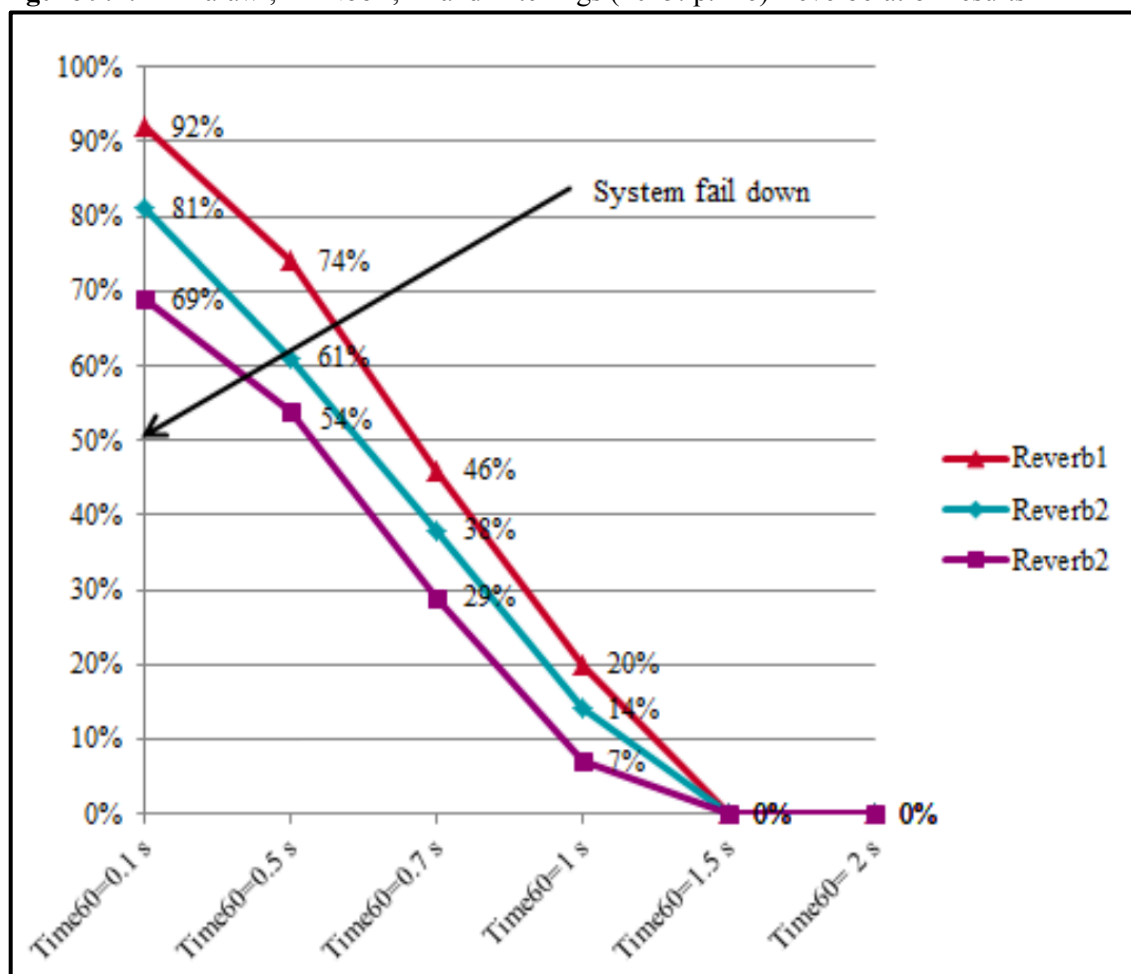
(GMM) become much closer together. Within this chapter a question was therefore set to explore whether i-vector statistical modelling ASRs would be more resilient.

Al-Karawi, Al-Noori, Li and Ritchings (2015) completed research experiments on the independent influence of noise (see chapter 8 on SNR) and reverberation. They used a Microsoft Speaker Recognition (MSR) tool kit to examine ASR performance. The toolkit can apply either GMM-UBM or i-vector processes.. For their research only GMM-UBM was selected. The team recorded 19 speakers (11 males and 8 females between 25 and 40 years old) at 16kHz sample rate (i.e. 0-8kHz frequency bandwidth). The speech recorded for the reverberation tests was text dependent and samples timed at between 30 and 40 seconds. They simulated reverberation using Matlab at (RT60) = 0.1s, 0.5s, 1.0s, 1.5s and 2s and documented meta-data pertaining to the complexity of reverberation (table 9.6). Their results are presented in % accuracy rather than EER (Figure 9.7). Baseline (i.e. clean) provided 100% accuracy. The term ‘system fail down’ was not fully explained - but it is inferred that this meant chance level accuracy. Normative data is not referred to. The extremely poor performance of the MSR system was noted on the relatively small reverberation settings and this provided further motivation for the experiments.

**Table 9.6:** Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) Reverberation settings

TABLE II: REVERBERATION SPECIFICATIONS			
Specification	Reverberation Model		
	Reverb 1	Reverb 2	Reverb 3
Room dimensions	3× 4×2.5 m	4× 4× 2.5 m	5× 4×2. 5 m
Romm volume	30 m <sup>3</sup>	40 m <sup>3</sup>	50 m <sup>3</sup>
Mic. Position	1	1.5	2 m
Source position	Fixed		
RT <sub>60</sub>	0.1 , 0.5 , 1 , 1.5 , 2 Second		
Walls reflection coefficients.	0.5 , 0.6 , 0.1 , 0.8		
Ceiling reflection coefficients.	0.9	0.9	0.9
Carpeted floor reflection coefficients.	0.6	0.6	0.6

**Figure 9.7:** Al-Karawi, Al-Noori, Li and Ritchings (2015: p.426) Reverberation results



Finally, with the increase in online videos and social media, it is likely that speech presented for forensic comparison could have passed through post processing plug-ins, such as reverberation. Websites such as [Waves.com](http://Waves.com) and [iZotope.com](http://iZotope.com), in addition to the spectral analysis of [Youtube.com](http://Youtube.com) material to assess for degradation, assisted with informing the recreation of room spaces for the purpose of this thesis.

### 9.3 Questions and Hypotheses

In reference to previous research and consistent with the core objectives in this thesis three questions were formed with the following associated hypotheses:

**Q1** How resilient are modern i-vector ASR systems to reverberation as opposed to the earlier GMM-UBM versions used in studies such as Castellano (1996) and Peer, Rafaely and Zigel (2008)? Further, how effective are session changes to an i-vector ASR system, based on adapting the normative data (UBM, TV, LDA+PLDA), relative to one another?

**H1** It is acknowledged that almost all ASR systems are inherently different in terms of configuration and settings, in addition to underlying architectural and normative session changes (GMM-UBM and i-vector/UBM, TV, LDA+PLDA). Nevertheless, performance will be assessed

relative to baseline (non-reverberant) data and it is suggested that i-vector systems should outperform GMM-UBM systems. This is due to the improvements in statistical modelling and the richness of the UBM, TV, LDA+PLDA normative data in comparison to GMM-UBM. This would also be in line with current research findings - e.g. Avila et al. (2015) (Figure 9.5).

For the second part of this question it is posited that an increase in the size, quality and relevance of the PLDA should initiate better ASR performance in EER%.

**Q2 Under a given set of conditions, can we quantify the influence of reverberation on ASR performance? If so, are there any direct correlations with specific reverberation measurements such as RT60?**

**H2** It is suggested that all reverberant conditions will have some detrimental effect on ASR performance but that quantifiable, direct mathematical correlations, will be difficult to extrapolate from limited data. However, it is hypothesised that an increase in RT60 will broadly result in poorer ASR performance. Larger rooms have longer RT60 values and a tendency towards generating greater complexity of reflections.

**Q3 Can the influence of reverberation be mitigated through:**

- **Matching conditions, i.e. RT60, for speaker model and test audio?;**
- **Adaptation or improvements to the normative data (i-vector/PLDA system) to potentially restore ASR performance?**

**H3** It is suggested that rectifying steps or processes applied once speech files have been affected by reverberant noise will offer no or very marginal benefit. However, experiments completed suggest that matching conditions could benefit performance and that amending the normative data (additional data) could also provide gains.

- Irrespective of the size of the room where the speaker models and test audio are matched it is likely there would be less detriment to the performance of the system than where they were unmatched.
- Improving normative data and adapting the feature extraction settings should partially restore performance.

### **9.3.1 Additional Experiment**

During the process of running the experiments it was determined that speech detection (VAD) was likely to be influencing the results from the i-vector/UBM, TV, LDA+PLDA system for the large environments with high RT60. A further experiment was run with VAD set to off in order to examine this in closer detail.



## 9.4 Methodology

When discussing the treatment of baseline data with reverberation it is important to differentiate the application of a processing plug in to recorded audio from re-recording the baseline material in a treated room. Both solutions could be regarded as artificial. In addition, it could not be assumed that the baseline data is totally devoid of all reverberation, since the interviews were conducted in a room rather than an anechoic chamber. It is conceded that an extremely small quantity of room reverberation was likely present in the baseline recordings, since avoidance would require anechoic recording. Nevertheless, on closer examination of spectrograms, the interview room clearly provides no audible reverberation (noticeable decay or ‘tail’) and the proximity to the microphone is very close, having been carefully set by a recording engineer. RT60 is estimated at almost zero - on the spectrogram it was unperceivable and could not be practically measured (i.e. less than .001 seconds). Reverberation was determined as negligible and deemed to be consistent throughout all the baseline recordings (same room, identical recording configuration). Therefore, ASR baseline performance was set for both GMM-UBM and i-vector/UBM, TV, LDA+PLDA systems from the recordings as supplied.

Two ASR systems were used. These were the OWR Vocalise 1 GMM-UBM system version 1.5.0.1190 and the OWR iVocalise i-vector system version 2.4.0.1547 (see Appendix G for full specifications). Bio-Metrics version 1.8.0.704 was used to chart, graph and plot performance results.

### 9.4.1 Reverberation Modeling

It was clearly not practical to re-record over 23 thousand audio files to test various different conditions. It was determined that artificially modelled reverberation would be the only practical approach, with batch processing used to model the large quantities of files required. The advantage of this approach is the relative consistency of the treatment to each file, although it could be argued that synthetic reverberation cannot truly match real world conditions. A very high-quality convolution reverberation plug-in was chosen to accurately and consistently model the complex reflections and absorption at different audio frequencies and for varying values of RT-60. Impulse Response Lite (IR-L) by [Waves.com](http://Waves.com) (2017) was selected, this being a respected industry standard. Additional information, provided by Waves, documented the significant lengths taken to accurately model the environments, including direct and early sound, through IR capture (see Appendix F). To broaden the library of settings, additional impulse responses were downloaded from the Waves website and assessed for suitability. Very large venue settings, such as Wembley Arena and Sydney Opera House, were not used as they were unlikely to be encountered in forensic casework. Settings were chosen to simulate more realistic casework conditions such as rooms in domestic properties and vehicles. A control test was conducted with the reverberation settings at zero or ‘dry’. This

was to establish baseline EER% and to ensure that the batch processing techniques, or any other part of the process, did not further alter the audio. Impulse response settings are summarized in Table 9.8.

**Table 9.8:** IR-L Reverberation settings selected for the experiments

No.	Reverberation setting (IR-L)	Convolution (seconds)	RT60 (seconds)	Dimensions (meters)	Distance (meters)
0	Control Test [Dry]	Baseline audio recordings (negligible)			
1	Living Room	1.92	0.7	6 x 4.6	3.0
2	Small Room	2.51	0.2	9 x 16	6.0
3	Kitchen	1.7	0.4	5 x 5.3	3.0
4	Bathroom	1.85	0.5	2.1 x 2.5	1.0
5	'Bluebird' Cafe	1.55	0.3	12 x 9	6.0
6	Lincoln Navigator Car	2.69	0.1	2.9 x 1.5	2.0
7	Ford Econoline Van 150	4.66	0.6	3.9 x 1.8	3.0
8	Bus	2.69	0.3	10 x 4.3	2.0
9	Hall	1.85	1.4	41.1 x 17.8	13.0
10	V Large Hall/Barbican	3.09	1.6	28.4 x 42.4	13.0

The plug-in treatments were applied to the baseline data and analysis was completed (spectrogram and auditory). Tiled surfaces, such as the kitchen and bathroom environments with harder surfaces, understandably produced more complex reflections and larger environments obviously produced greater values of RT60. Conversely smaller interiors with more soft furnishings, such as the vehicles, provide more absorption and lower RT60. Microphone proximity (relating to the impulse response modelling) was also documented within the supporting Waves documentation, as the complexity of reverberation reflections is influenced by distance between capture and reflective surfaces. Perhaps unrealistically, there were also few people physically present, again providing less absorption. Finally, the proximity of the speaker to the microphone remained constant across all DyViS recordings and therefore the modelled material is more consistent with scenarios involving static speakers than those moving around a room, for example.

## 9.4.2 Data Preparation

Baseline data from the task 1 interview DyViS data was created. This consisted of 100 male speakers SSBE. The files were edited to create 100 speaker models (1m per speaker) x 300 test audio files 1m, 1m, residual to provide 30,000 cross comparisons. 10 reverberation settings were applied using batch processing. As discussed, this was completed with the Waves IR-L plug in and Reaper ([Reaper.fm](http://Reaper.fm), 2017) to batch process. 10 new sets of speaker models and test audio files were created. A data set without reverberation applied (referred to as 'dry') was set aside to validate the batch processing output.

Izotope RX6 Advanced ([iZotope.com](http://izotope.com)) was used to examine the consistency of output for the 10 environments, using a combination of spectrograms and auditory analysis. Initially the multi-way cross comparisons were processed through iVocalise with the same PLDA settings as the baseline tests. Results were examined using OWR Bio-Metrics software (EER%, DET, LR Plot, Zoo plots), Excel and Izotope RX6 Advanced.

#### **Experiments set:**

- i. A control test for each ASR (GMM-UBM and i-vector) on non-processed baseline audio files (100xSM compared to 300xTA) with all reverberation settings on bypass/dry.
- ii. Matched conditions: 10 batches of speaker models with reverberation applied (100 speakers) compared against 10 batches of test audio files with the same reverberation applied (3 files per speaker) for both ASR systems.
- iii. Unmatched conditions: 10 batches of speaker models, with no reverberation applied (100 speakers) compared against 10 batches of test audio files with varying degrees of reverberation applied (3 files per speaker) for both ASR systems.
- iv. Finally, two additional iVocalise sessions set (UBM, TV, LDA+PLDA) were constructed and trained in consultation with OWR to test the influence of normative data on i-vector ASR performance (please see 9.5.1). All i-vector experiments from session 1 were re-run, for matched and unmatched conditions in all environments, using the additional bespoke UBM, TV, LDA+PLDA sets labelled ‘PLDA 2 and 3’.

### **9.4.3 Normative Data, Gaussian Mixture Model System**

A bespoke UBM was created for the GMM-UBM system to account for the interview channel data. Eighty-nine speakers were selected from the Speech Obtained in Key Environments corpora and database (Alexander et al. 2015), or SPOKE, interview data. The MFCC extraction settings were adapted to account for the increased frequency bandwidth representative of the DyViS interview data (16kHz sample rate/0-8kHz frequency bandwidth).

### **9.4.4 Normative Data, i-vector System**

The following three UBM, TV, LDA+PLDA sessions were built specifically for this set of research experiments. As part of the requirement, all normative sets did not contain any DyViS material and were optimised for the purpose of wide band/interview speech.

This section provides a brief technical description of the contributing data and MFCC criteria used to create the three UBM, TV, LDA+PLDA sessions used in this chapter. It was informed by technical correspondence from OWR (Dr. Anil Alexander and Dr. Finnian Kelly).

### **PLDA Session 1: SREPRISM 27k 2016C TEL 1024DD-AnilBuild-NoDyvis.xmlsession**

The total number of speech files used in training this PLDA session is 37,960. There are approximately 27,150 NIST<sup>§§</sup> SRE files of around 3 minutes per file. This corresponds to approximately 1,357 hours. This session uses an extraction process set at 13 MFCC features, with deltas and delta-deltas.

### **PLDA Session 2: MEGA PRISM 38k TEL PLUS-2016C-NoDYVIS-WithoutSITW-13DDCMS-AnilBuild.xmlsession**

This session contains the NIST SRE files as listed in PLDA1 with additional LDC<sup>\*\*\*</sup> data. This provided additional speech files recorded in interview settings (i.e. small/low reverberant spaces). The number of additional reverberant files used and the sizes of recording spaces was not specified due to lack of metadata, captured at the recoding stage and so could not be supplied. The total number of files used in training in both these session is 46,673 files. The combined session contains approximately 37,888 files each of around 3 minutes duration. This provides approximately 1,894 hours of audio. This session uses an extraction process set 13 MFCC features with only deltas.

### **PLDA Session 3: MEGA PRISM 38k TEL PLUS-2016C-NoDYVIS-WithoutSITW-13DDCMS-AnilBuild-Mk2**

This session uses all the speech files as listed in PLDA2 and only differs in using a different feature extraction process (13 MFCC features in addition to delta and delta-delta features, as with PLDA1).

## **9.5 Results**

**Table 9.9:** GMM-UBM Results. Matched conditions  
Vocalise 1 GMM-UBM. Matched SM and TA conditions.

<b>Reverb Type</b>	<b>RT 60</b>	<b>EER %</b>	<b>Cllr</b>	<b>H0 Mean</b>	<b>H1 Mean</b>	<b>H0 SD</b>	<b>H1 SD</b>	<b>FAR, FRR % per 100</b>	<b>FAR, FRR % per 1,000</b>	<b>FAR, FRR % per 10,000</b>
<b>Control</b>	N/A	7.56	2.25	5.84	3.04	0.96	1	33.69	60.67	77.33
<b>L/Room</b>	0.7	10.08	1.29	2.8	1.5	0.51	0.55	52.22	79	88.67
<b>S/Room</b>	0.2	7.7	1.42	3.4	1.73	0.59	0.64	40.24	73	82.51
<b>Kitchen</b>	0.4	8.36	1.51	3.42	1.9	0.56	0.62	44.79	76.28	84.33
<b>B/room</b>	0.5	8.03	1.32	2.96	1.56	0.5	0.56	46.33	73.81	82.35
<b>Cafe</b>	0.3	8.36	1.38	3.24	1.66	0.57	0.63	42.33	72.1	83.33
<b>Car</b>	0.1	8.29	1.94	4.78	2.58	0.77	0.84	36.93	68.67	77.68
<b>Van</b>	0.6	5.7	1.7	4.61	2.21	0.69	0.83	28.27	58.1	71.36
<b>Bus</b>	0.3	8.66	1.52	3.59	1.92	0.61	0.66	42.67	72.67	86.01
<b>Hall</b>	1.4	9.43	1.08	2.08	1.05	0.4	0.44	51.33	74.48	92.67
<b>L. Hall</b>	1.6	10.54	1.07	2.04	1.03	0.39	0.47	52.04	79.88	92.68

<sup>§§</sup> National Institute of Standards and Technology [nist.gov/](http://nist.gov/) Speaker Recognition Evaluation

<sup>\*\*\*</sup> Linguistic Data Consortium [ldc.upenn.edu/](http://ldc.upenn.edu/)

**Table 9.10:** GMM-UBM Results. Unmatched conditions  
Vocalise 1 GMM-UBM. Unmatched SM and TA conditions.

Reverb Type	RT 60	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR % per 1,000	FAR, FRR % per 10,000
Control	N/A	7.56	2.25	5.84	3.04	0.96	1	33.69	60.67	77.33
L/Room	0.7	12.9	0.74	1.43	0.08	0.58	0.7	68.83	90.22	98
S/Room	0.2	8.91	1.12	2.9	1.17	0.65	0.72	51.33	72.33	84.33
Kitchen	0.4	8.67	1.12	2.84	1.17	0.61	0.7	51.33	74.61	86
B/room	0.5	11.35	0.89	2.09	0.61	0.59	0.69	61.33	83.43	89.34
Cafe	0.3	9.14	1.13	2.92	1.97	0.63	0.71	50.33	74.43	85.34
Car	0.1	8.65	1.83	4.7	2.4	0.81	0.88	40.08	67.05	75.51
Van	0.6	7.78	1.6	4.33	2.04	0.8	0.87	41.56	65.1	78.67
Bus	0.3	9.31	1.28	3.3	1.48	0.68	0.75	48.17	72.97	82.68
Hall	1.4	21.4	0.76	0.2	-0.8	0.58	0.69	81	95	98.33
L. Hall	1.6	27.16	1.11	-1.01	-1.92	0.68	0.86	91	96.67	98.33

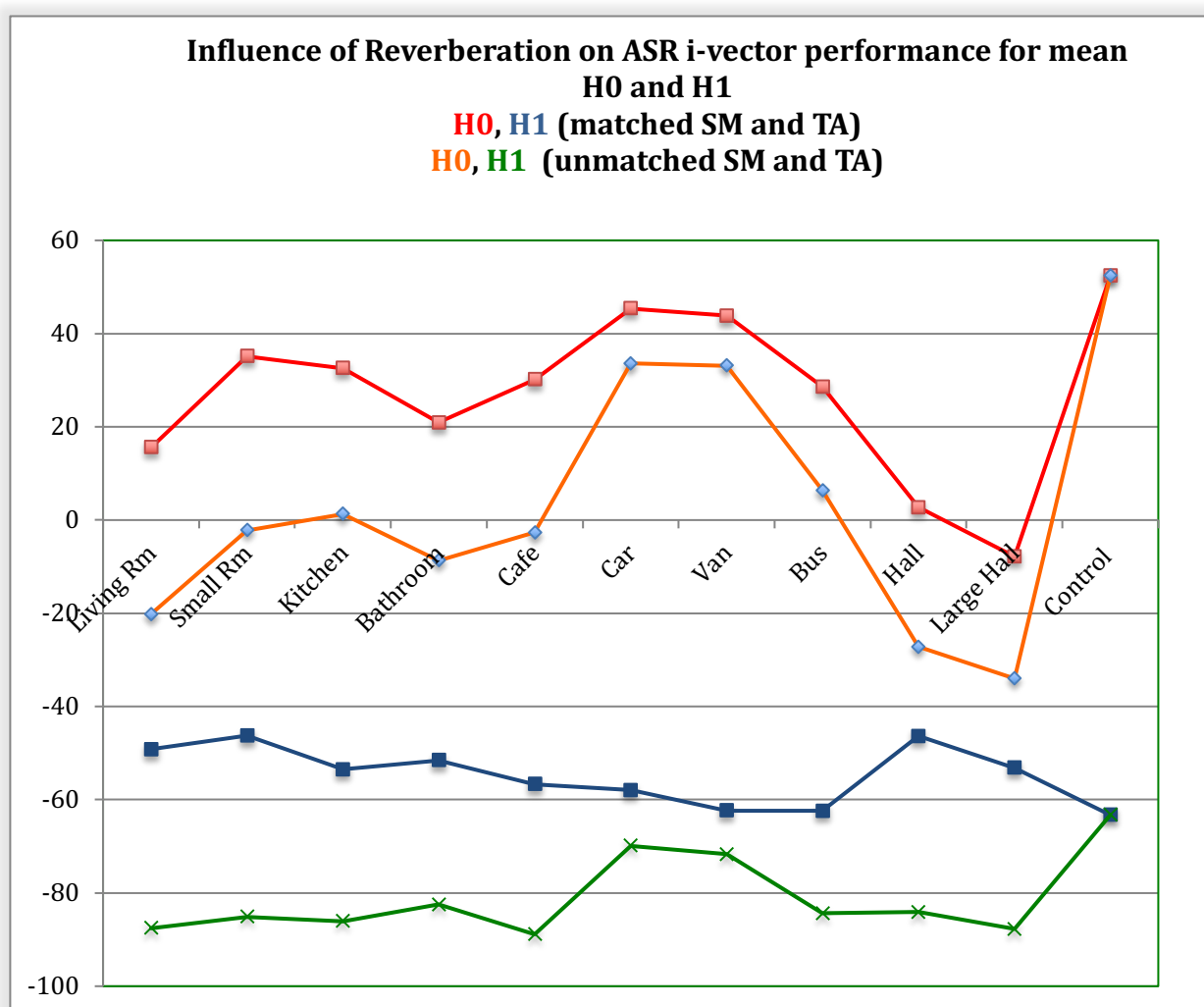
**Table 9.11:** I-vector/UBM, TV, LDA+PLDA results. Matched conditions  
Matched SM and TA with bespoke PLDA session 1

Reverb Type	RT 60	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR % per 1,000	FAR, FRR % per 10,000
Control	N/A	0.05	0.02	52.37	-63.27	12.31	27.46	0.00	0.10	1.00
L/Room	0.7	4.24	1.46	15.56	-49.17	16.76	21.10	13.57	29.10	56.01
S/Room	0.2	1.04	0.74	35.13	-46.24	13.39	23.35	1.33	8.00	21.01
Kitchen	0.4	1.01	0.12	32.59	-53.50	14.16	23.25	1.00	5.77	19.69
B/room	0.5	2.77	0.59	20.92	-51.54	15.49	22.26	5.70	19.80	32.57
Cafe	0.3	0.99	0.16	30.17	-56.69	14.25	23.64	1.00	6.15	15.00
Car	0.1	0.33	0.03	45.41	-57.92	13.04	25.71	0.00	1.00	3.68
Van	0.6	0.38	0.03	43.87	-62.33	12.89	25.64	0.00	1.00	5.67
Bus	0.3	1.01	0.17	28.59	-62.40	13.84	24.08	1.33	4.00	8.01
Hall	1.4	10.29	4.64	2.73	-46.36	18.71	20.07	38.77	61.44	78.67
L. Hall	1.6	14.28	8.83	-7.83	-53.14	19.59	19.73	42.77	66.48	85.35

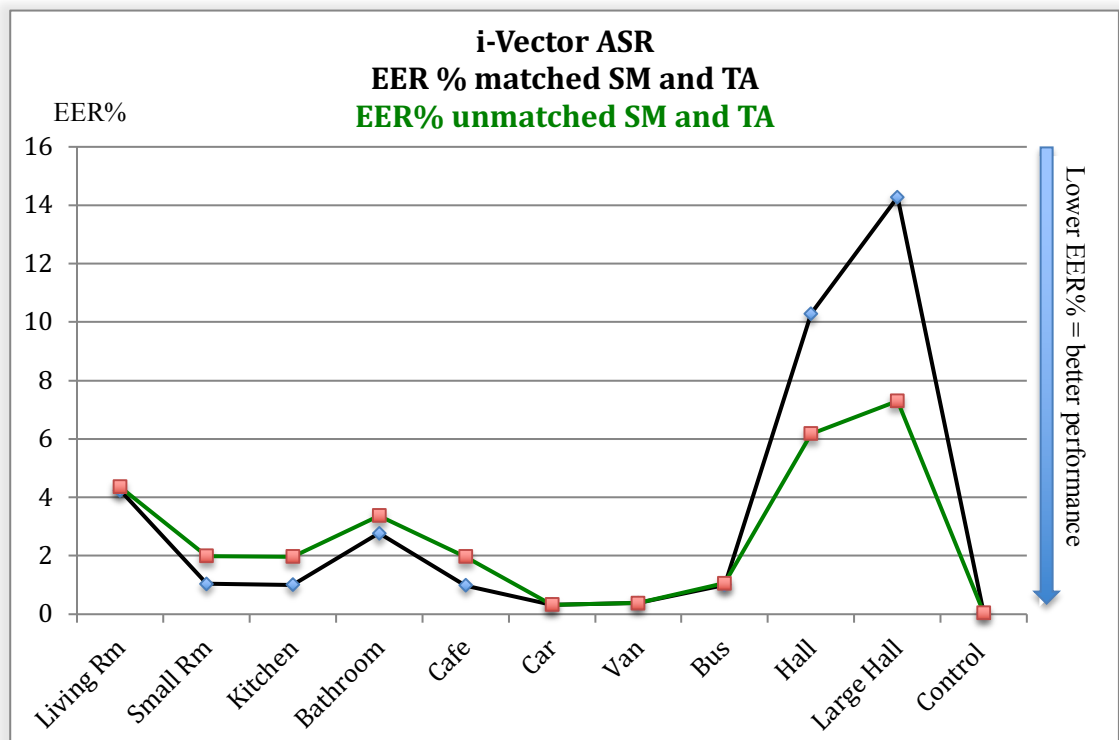
**Table 9.12:** I-vector/UBM, TV, LDA+PLDA results. Unmatched conditions  
Unmatched SM & TA with bespoke PLDA session 1

Reverb Type	RT 60	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR FRR % per 1,000	FA RF RR % per 10,000
Control	N/A	0.05	0.02	52.37	-63.27	12.31	27.46	0.00	0.10	1.00
L/Room	0.7	4.36	15.39	-20.21	-87.57	17.79	22.35	17.33	47.07	74.03
S/Room	0.2	1.99	5.25	-2.23	-85.14	15.45	24.19	3.00	19.1	61.36
Kitchen	0.4	1.97	4.32	1.32	-86.04	16.66	24.34	3.67	14.67	39.72
B/room	0.5	3.37	8.36	-8.65	-82.47	16.24	23.28	9.06	27.26	63.79
Cafe	0.3	1.97	5.85	-2.67	-88.87	17.14	24.3	4.11	15.77	41.36
Car	0.1	0.33	0.70	33.57	-69.91	13.17	26.6	0.00	1.33	8.68
Van	0.6	0.38	0.11	33.12	-71.68	12.85	26.25	0.00	1.67	5.00
Bus	0.3	1.06	2.60	6.39	-84.39	15.51	25.05	1.33	12.33	32.36
Hall	1.4	6.18	19.96	-27.23	-84.12	17.7	20.9	24.62	64.62	72.86
L. Hall	1.6	7.3	24.56	-34	-87.77	17.94	21.68	44.19	67.57	86.49

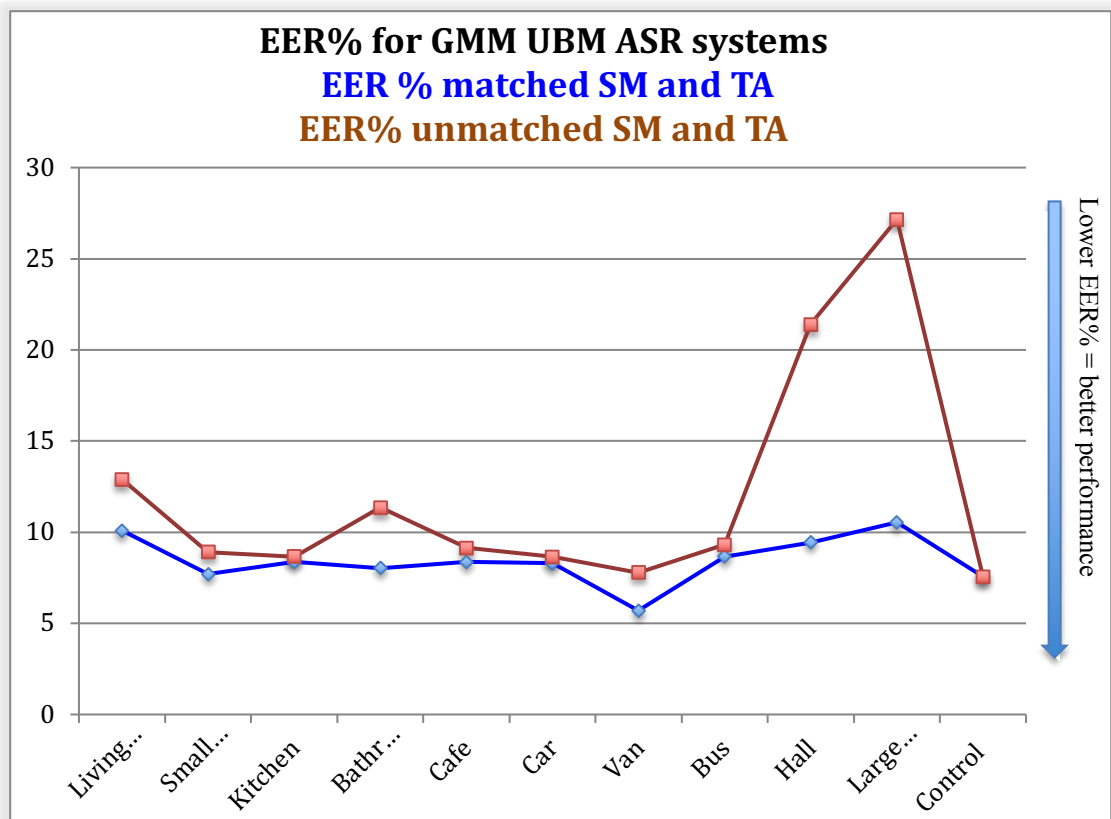
**Figure 9.13:** Influence of reverberation on i-vector/UBM, TV, LDA+PLDA ASR H0 and H1 for matched and unmatched conditions, SM and TA.



**Figure 9.14:** Influence of reverberation on i-vector/UBM, TV, LDA+PLDA EER% I-vector EER% for matched and unmatched conditions, SM & TA.



**Figure 9.15:** Influence of reverberation on GMM-UBM ASR EER% GMM-UBM for matched and unmatched conditions, SM & TA.



### 9.5.1 Observations

In line with predictions and consistent with previous research, ASR performance decreased as the complexity and size of the reverberation rose (RT60). This was consistent for both GMM-UBM and i-vector/UBM, TV, LDA+PLDA systems. However, compared with the results from the GMM-UBM ASR the i-vector system fared more favourably and appeared to demonstrate a smaller performance decrease when presented with light to moderately reverberant material and unmatched SM and TA (Figures 9.14 and 9.15). It should also be considered that the technical quality of the CTEST/SPOKE interviews (UBM) was marginally lower overall in comparison to DyViS and the number of speakers small (89) and normative sets for i-vector systems are significantly larger by design. It could be argued that a larger UBM for the GMM system, of higher quality, could marginally improve performance. Indeed, further experiments on baseline data were conducted on the GMM UBM system with various normative changes and settings adjusted. However, whilst the performance fluctuated marginally, including a marginally improved EER of 3.018% if applying band limiting on file ingest (0-4kHz frequency bandwidth) despite many adjustments, the performance of the i-vector system was consistently and considerably better than the GMM-UBM system even given the inherent architectural differences.

Score separation (distance between same speaker and different speaker distributions) was marginally improved under matched SM and TA conditions as opposed to unmatched for both systems. Rooms with relatively low values of RT60 but poor absorption, such as the tiled bathroom, exhibited performance degradation with higher EER% than predicted. This is likely due to less absorption and multiple/complex reflections i.e. highly reflected sound waves, not absorbed by surfaces/furniture, merging together at high speed which then interferes with sub-second frame measurements (of the ASR system). Conversely, the van with larger RT60 values provided better EER% performance than predicted, likely due to greater absorption and lower complexity of reflections.

The i-vector ASR system was much more robust than expected against reverberation degradation. Only marginal drops in EER% and Cllr performance were noted for the car and van modelled environments, for example. This is likely due to the dampening of reflections caused by the sound absorbing materials inside the vehicles, such as the seat, carpets, roof linings etc. In contrast, and as predicted, reverberation settings with larger RT60 times and/or longer convolution times tended to degrade ASR performance much more significantly. This particularly applied to the much larger spaces (living room, hall and large hall).

As predicted the matched conditions performed better than unmatched conditions (i.e. degradation of both speaker models and test audio files). However, there were exceptions for the two largest reverberation settings (hall and large hall). The GMM-UBM system performed marginally better

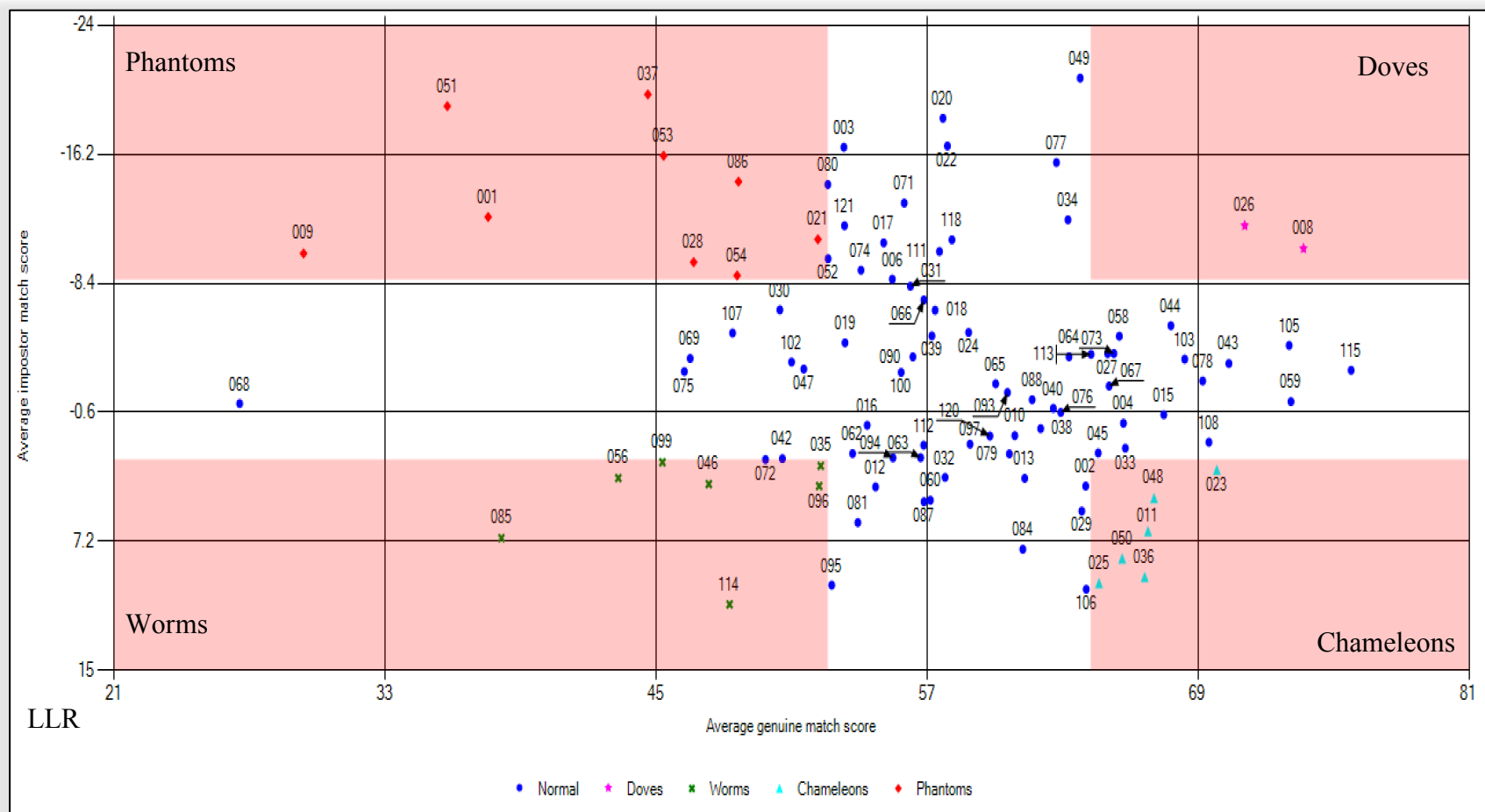


for those two environments (matched) and it was not known why – although one plausible explanation could be that broader, rather than more detailed, statistical modelling may be of benefit in the GMM-UBM system. Alternatively, the quantity of non-degraded speech passing through the voice activity detection algorithm embedded in the feature extraction stage (VAD) in the i-vector system for the untreated speaker models may have been simply less. On closer examination, it was observed that the quantity of net speech extracted by the VAD fell sharply for the long and complex reverberation treatments on the i-vector system and this was most noticeable for the speaker models, with some reducing to as little as 10s in duration. Conversely, when SM and TA were unmatched, more speech passed through to the feature extraction for the speaker models resulting in subsequent performance improvements. More research is required.

It was also evident, from the H0, H1 graphs, that without further normalisation or calibration it would be extremely difficult in casework to interpret LR results based on score height alone and we have shown that this can be influenced by reverberation. To examine this further, the ASR output was also analysed using zoo plots. Several additional observations were noted for ASR scores that passed through the reverberation process when compared to the control data and an example is presented below (Figure 9.16 and Figure 9.17).



**Figure 9.17:** Zoo plot of Living Room results, matched conditions, i-vector, PLDA 3



Note: the axis numbering and scales of the two figures above are different, due to the wide variation in distribution between results.

## Zoo plot observations

The zoo plots clearly show ASR performance degradation of the reverberant data against baseline.

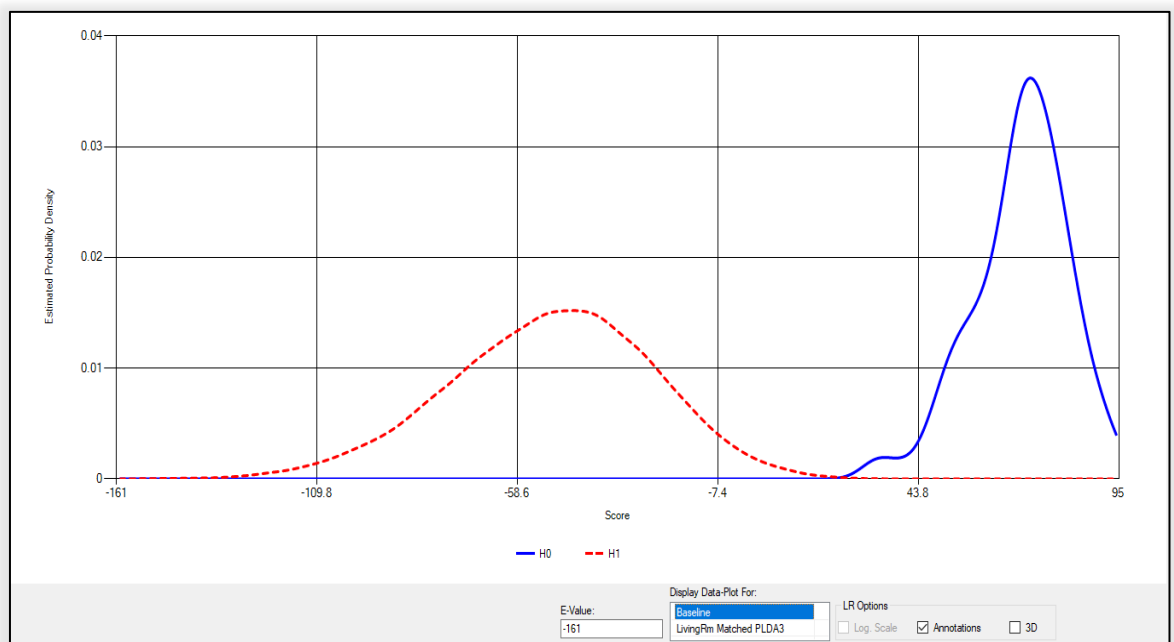
- i. Fewer Doves (speakers more easily verified by the ASR with high match scores and low non-match scores) were noted in the reverberant data results in comparison with the baseline results. This further declined relative to higher RT60 values.
- ii. Conversely, a relatively large increase of problematic speaker categories (i.e. Worms, Phantoms and Chameleons) was noted in the reverberant results with greater numbers relative to higher RT60 values.
- iii. Overall, a typical trend of plot distribution movement from upper right to lower left was observed with greater dispersal.

For practical implementation, this equates to greater ASR speaker confusability of reverberant audio over non-reverberant material. It is possible that it could be compensated for (i.e. through improving calibration, augmenting PLDA training data and/or threshold settings). However, it is strongly recommended that experienced interpretation of output and examination of auditory phonetic content supports acoustic results – although it is conceded that this might not be possible either, dependent on the extent of the reverberation.

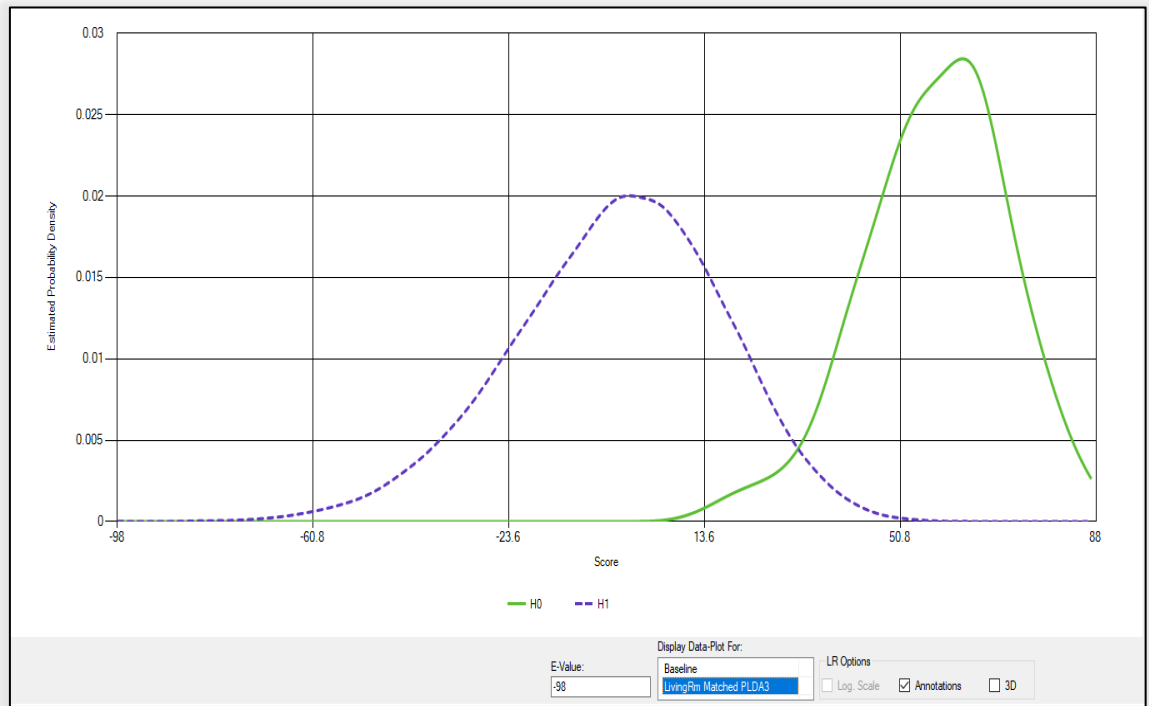
The ASR output was also examined using LR plots. Observations were noted for ASR scores that passed through the reverberation process, when compared to the baseline data.

## LR plot observations

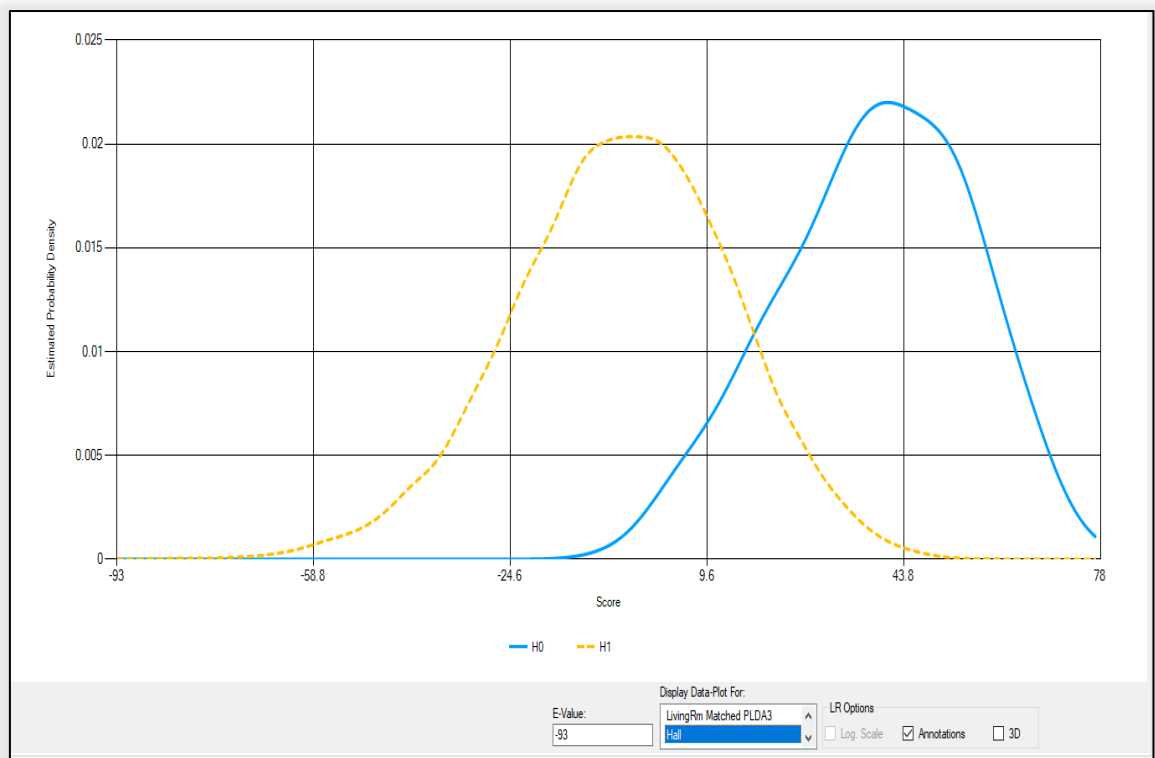
**Figure 9.18:** LR Plot, Baseline data, matched conditions, PLDA session 3



**Figure 9.19:** LR Plot, Living Room data, matched conditions, PLDA session 3



**Figure 9.20:** LR Plot, Hall data, matched conditions, PLDA session 3



The following observations were noted from the LR plots. Note the variation in axis scales re LR across Figures 9.18, 9.19 and 9.20. In addition:

- i. Lower same speaker scores and higher different speaker LR scores were observed for the reverberant data as RT60 increased (e.g. living room and hall conditions in Figures 9.19 and 9.20) even for optimum ASR settings (PLDA3);

- ii. Discrimination degradation (i.e. lower same speaker and higher different speaker scores) occurred, broadly as RT60 values increased;
- iii. Decline noted in same speaker score height with wider spread of score distribution noted (broadening of bell-curve);
- iv. Same speaker and different speaker distribution curves began to merge as RT60 increased. This highlights the difficulties with regard to setting thresholds as greater overlapping between distributions occurs.

## **9.5.2 System Accuracy Results**

The OWR Bio-Metrics software generates Cllr scores based on a standard calculation (Brummer and D. Van Leeuwen, 2006) and this was applied for checking system accuracy. As discussed (chapter 3) a lower Cllr value indicates a more accurate and precise system with Cllr <1 widely viewed as an acceptable level of accuracy (3.5.5).

Matched and unmatched conditions produced similar EER%, but with different Cllr (Table 9.21).

**Table 9.21:** I-vector ASR tests (PLDA session 1). Examination of Cllr

<b>Reverb Type</b>	<b>Matched (SM/TA) or Unmatched</b>	<b>RT60</b>	<b>EER%</b>	<b>Cllr</b>
Control	-	N/A	0.05	0.02
L/Room	Matched	0.7	4.24	1.46
L/Room	Unmatched	0.7	4.36	15.39
S/Room	Matched	0.2	1.04	0.74
S/Room	Unmatched	0.2	1.99	5.25
Kitchen	Matched	0.4	1.01	0.12
Kitchen	Unmatched	0.4	1.97	4.32
B/room	Matched	0.5	2.77	0.59
B/room	Unmatched	0.5	3.37	8.36
Cafe	Matched	0.3	0.99	0.16
Cafe	Unmatched	0.3	1.97	5.85
Car	Matched	0.1	0.33	0.03
Car	Unmatched	0.1	0.33	0.70
Van	Matched	0.6	0.38	0.03
Van	Unmatched	0.6	0.38	0.11
Bus	Matched	0.3	1.01	0.17
Bus	Unmatched	0.3	1.06	2.60
Hall	Matched	1.4	10.29	4.64
Hall	Unmatched	1.4	6.18	19.96
L. Hall	Matched	1.6	14.28	8.83
L. Hall	Unmatched	1.6	7.3	24.56

Under matched conditions using the i-vector/PLDA system the Cllr was consistently lower than for unmatched conditions. Of course, results do not take into consideration any system calibration – which is unlikely to alter EER% (discrimination) but can influence Cllr (accuracy). As previously stated, calibration is specifically not applied in the experiments so as not to conflate variables. In the experiments completed, matched conditions likely provide a naturally calibrated system (assuming normative data is relevant to conditions). It was also noted that van and car environments provided relatively small decreases in Cllr, demonstrating relative (accuracy) resilience to light reverberation. It was also observed that in unmatched conditions Cllr values rose considerably in line with reverberation increase (RT60 and complexity as captured and modelled by IR).

Finally, it will be shown (9.5.3), that Cllr also elevated in all cases as the session/PLDA data size grew. This suggests that, whilst there is some benefit in raising the quantity of material in the PLDA, there is a point of diminishing returns at which more data makes actually negligible further difference to system EER% and can actually decrease accuracy. Further research is suggested to determine optimum size of normative data.

### 9.5.3 Results from Normative Sessions 2 and 3

As discussed, in consultation with OWR a second and third bespoke normative set (UBM, TV, LDA+PLDA) was constructed (see 9.4.4). For ease of reference, the initial normative set is defined as PLDA version 1 and the subsequent tests are referenced as PLDA 2 and 3. As set out in the research questions the objective was to test the hypothesis that improvements could be achieved through:

- i. Increasing the PLDA data (population size) with additional speech corpora (NIST and LDC).
- ii. Adding reverberant recordings into the normative data/PLDA.

Experiments were re-run, utilising the new PLDAs (versions 2 and 3) for both matched and unmatched conditions. Results are presented in Tables 9.22 to 9.26.

#### PLDA 2 Results

**Table 9.22:** Summary results from reverberation experiments PLDA2, matched **Matched SM and TA Bespoke PLDA2**

Reverb Type	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1,000	FAR, FRR 10,000
Living Rm	2.97	3.09	53.22	-9.30	12.97	20.82	7.22	22.67	42.00
Small Rm	1.04	2.79	64.47	-11.03	10.79	21.36	1.33	4.40	30.02
Kitchen	1.30	3.00	65.47	-9.20	11.47	20.28	1.33	5.67	15.33
Bathroom	1.35	1.15	50.81	-19.53	11.44	21.15	1.68	7.09	21.14
Cafe	0.38	0.66	56.80	-27.01	11.53	22.30	0.00	1.67	7.67
Car	0.08	0.40	67.87	-36.01	12.27	24.85	0.00	0.33	3.00
Van	0.14	0.18	65.62	-44.17	12.08	25.44	0.00	0.00	4.33
Bus	0.31	0.35	57.29	-35.05	12.13	23.97	0.00	2.00	9.00
Hall	11.32	3.79	39.55	-6.60	17.09	20.26	41.33	66.33	88.33
L. Hall	12.04	1.83	26.82	-18.78	18.09	20.05	41.56	64.98	86.67
Control	0.007	0.08	68.66	-55.11	12.15	27.50	0.00	0.00	0.33

**Green** = improvement on previous PLDA EER% outcomes.

**Red** = poorer than previous PLDA EER% outcomes.



**Table 9.23:** Summary results from reverberation experiments PLDA2, unmatched  
**Unmatched SM and TA**

Reverb Type	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1,000	FAR, FRR 10,000
Living Rm	4.06	15.58	-21.00	-83.37	15.89	20.71	11.29	31.33	52.67
Small Rm	1.00	2.23	6.08	-72.06	13.68	22.42	1.00	13.33	56.36
Kitchen	1.67	2.05	7.86	-73.67	14.42	22.26	2.33	9.55	37.68
Bathroom	2.08	5.17	-3.14	-74.98	14.49	21.74	5.70	21.70	52.37
Cafe	1.04	2.7	6.32	-78.54	15.45	23.15	1.67	11.77	34.33
Car	0.05	0.04	50.68	-57.39	11.97	25.98	0.00	0.00	2.34
Van	0.05	0.02	47.86	-62.42	12.40	26.02	0.00	0.33	5.00
Bus	1.27	0.76	17.85	-74.46	14.59	24.43	1.67	4.67	19.34
Hall	10.30	34.89	-48.37	-92.79	16.30	19.18	46.67	72.14	89.67
L. Hall	11.67	36.52	-50.63	-91.29	16.00	18.42	48.07	74.00	91.35
Control	0.007	0.08	68.66	-55.11	12.15	27.50	0.00	0.00	0.33

### PLDA 3 Results

**Table 9.24:** Summary results from reverberation experiments PLDA3, matched  
**Matched SM and TA**

Reverb Type	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1,000	FAR, FRR 10,000
Living Rm	3.68	4.70	57.42	-3.11	13.55	20.11	7.33	26.88	50.00
Small Rm	1.33	3.80	66.70	-6.16	10.67	20.30	1.67	4.00	22.02
Kitchen	1.00	4.62	69.02	-3.12	11.24	19.85	1.00	5.43	15.01
Bathroom	1.40	1.71	55.55	-15.92	12.17	21.07	2.01	9.14	24.50
Cafe	0.67	1.22	61.31	-21.32	11.71	22.48	0.00	3.70	17.67
Car	0.33	0.42	68.89	-34.70	12.02	24.23	0.00	0.33	1.67
Van	0.02	0.17	65.83	-43.26	12.16	24.52	0.00	0.00	3.01
Bus	0.34	0.65	61.37	-29.76	12.18	23.94	0.00	2.67	10.00
Hall	11.64	3.83	38.83	-5.16	16.44	19.13	41.67	65.77	80.67
Large Hall	11.73	1.79	24.80	-19.70	18.18	19.42	40.85	67.43	86.69
Control	0.005	0.11	69.92	-49.98	11.97	26.07	0.00	0.00	1.67

**Green** = further improvement on previous (PLDA 2) EER% outcomes

**Red** = poorer than previous (PLDA 2) EER% outcomes

**Table 9.25:** Summary results from reverberation experiments PLDA3, unmatched  
**Unmatched SM and TA, bespoke PLDA 3**

Reverb Type	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1,000	FAR, FRR 10,000
Living Rm	4.34	18.49	-25.22	-84.67	16.24	19.96	16.07	34.37	57.00
Small Rm	1.33	2.12	6.55	-68.57	14.00	21.78	2.00	16.10	59.02
Kitchen	1.67	2.27	6.39	-72.02	14.26	21.74	2.13	11.92	36.01
Bathroom	2.95	5.14	-2.60	-73.09	15.34	21.84	7.05	23.24	58.76
Cafe	1.60	3.61	4.08	-78.26	16.43	23.24	2.33	15.20	42.68
Car	0.03	0.05	52.55	-53.66	11.71	24.96	0.00	0.00	2.00
Van	0.33	0.018	47.74	-60.52	12.72	25.06	0.00	0.33	2.33
Bus	1.01	1.02	15.52	-74.47	15.22	24.15	1.00	7.33	30.33
Hall	13.17	35.52	-49.23	-90.61	17.43	18.97	49.52	76.43	92.68
Large Hall	13.73	36.89	-51.14	-89.55	16.27	18.27	54.23	79.77	93.01
Control	0.005	0.11	69.92	-49.98	11.97	26.07	0.00	0.00	1.67

In summary, performance benefit was demonstrated through improvements to the PLDA and this was particularly evident under matched conditions. The addition of more data and inclusion of reverberant material further enhanced performance. However, performance was not uniformly improved across all conditions with a single PLDA, particularly when considering Cllr (accuracy) in addition to EER% (Table 9.25).

**Table 9.26:** EER% Optimal performance across PLDA sessions 1, 2 and 3

Condition	PLDA 1 EER%	PLDA 2 EER%	PLDA 3 EER%
Living Room	4.24 (m)	2.97 (m)	3.68 (m)
Small Room	1.04 (m)	1.00 (u)	1.33(x)
Kitchen	1.01 (m)	1.30 (m)	1.00 (m)
Bathroom	2.77 (m)	1.35 (m)	1.44 (m)
Cafe	0.99 (m)	0.38 (m)	0.67 (m)
Car	0.33 (m)	0.05 (u)	0.03 (u)
Van	0.38 (m)	0.05 (u)	0.02 (m)
Bus	1.01 (m)	0.31 (m)	0.33 (u)
Hall	6.18 (u)	10.30 (u)	11.64 (m)
Large Hall	7.30(u)	11.67 (u)	11.73 (m)
Control	0.05	0.007 (x)	0.005 (x)

**Best overall performance in Green.**

**m = matched: u = unmatched.**

**x = Denotes identical results obtained for both matched and unmatched conditions (within .001 EER%)**

For the i-vector system tests, 66.7% of experiments excluding baseline performed better (or equal) under matched conditions. UBM, TV, LDA+PLDA relevance and size had more influence on performance than matched/unmatched conditions alone. However, although the baseline EER% consistently fell with the addition of more data this was not necessarily the case for reverberant material. Results demonstrated that increasing PLDA data provided better performance overall for those conditions for which reverberation was low (e.g. vehicles) but not necessarily for larger environments, where performance actually fell in some instances. PLDA 3 provided only a marginal improvement over PLDA2 in EER% in just a few conditions – likely suggesting data saturation/diminishing returns.

## 9.5.4 Speech Detection Results

As discussed (9.3.1) an additional test was run, using PLDA session 3 to determine the difference in results when deselecting the speech detection algorithm.

**Table 9.27:** VAD Results. Matched conditions, PLDA session 3, VAD Off

Reverb Type	RT 60	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1,000 %	FAR, FRR 10,000 %
<b>Control</b>	N/A	0.0017	0.15	74.35	-48.00	11.72	26.34	0.00	0.00	0.00
<b>L/Room</b>	0.70	2.05	9.73	71.80	9.26	10.82	19.40	3.00	10.00	27.00
<b>S/Room</b>	0.20	0.40	4.52	73.07	-3.98	9.76	20.42	0.21	2.33	7.34
<b>Kitchen</b>	0.40	0.75	6.55	76.41	1.91	9.79	20.03	0.67	2.33	4.17
<b>B/room</b>	0.50	0.66	3.33	66.57	-8.04	10.54	20.68	0.67	1.67	10.00
<b>Cafe</b>	0.30	0.37	1.52	67.74	-19.13	10.66	22.48	0.00	1.33	4.35
<b>Car</b>	0.10	0.31	0.49	73.54	-34.28	11.57	24.85	0.00	0.67	1.67
<b>Van</b>	0.60	0.01	0.21	70.76	-42.29	11.45	25.01	0.00	0.00	0.33
<b>Bus</b>	0.30	0.29	0.95	68.64	-25.83	10.89	23.94	0.00	0.67	2.34
<b>Hall</b>	1.40	3.67	12.3	62.95	15.09	11.10	17.02	15.00	38.38	58.69
<b>L. Hall</b>	1.60	6.33	5.16	50.12	-0.69	13.42	18.62	16.67	34.63	69.69

**Table 9.28:** VAD Results. Unmatched conditions, PLDA 3, VAD Off

Reverb Type	RT 60	EER %	Clr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR %	FAR, FRR 1000 %	FAR, FRR 10000 %
Control	N/A	.0017	0.15	74.35	-48.00	11.72	26.34	0.00	0.00	0.00
L/Room	0.70	3.99	16.79	-22.93	-84.55	15.21	20.28	10.83	30.67	51.68
S/Room	0.20	1.32	1.01	13.06	-67.70	14.38	22.53	1.33	11.94	57.33
Kitchen	0.40	1.66	1.60	11.10	-70.68	14.83	22.29	2.85	9.48	32.67
B/room	0.50	2.01	3.97	0.56	-74.22	14.96	22.21	3.24	19.70	53.35
Cafe	0.30	1.11	2.38	8.93	-78.89	16.63	23.51	2.00	11.67	37.00
Car	0.10	0.03	0.07	57.87	-52.21	11.79	25.40	0.00	0.00	2.00
Van	0.60	0.01	0.03	52.80	-59.04	12.69	25.56	0.00	0.00	2.33
Bus	0.30	0.84	0.51	21.04	-73.66	14.90	24.60	0.33	3.33	22.01
Hall	1.40	9.71	35.90	-49.77	-93.03	15.76	18.69	43.67	76.33	87.00
L. Hall	1.60	11.98	36.26	-50.26	-90.82	15.63	18.54	47.93	79.10	94.00

Some performance benefit was demonstrated by switching VAD off. This was evident for most reverberation settings with better EER% performance overall and higher accuracy (lower Clr). It is possible that this is due to the VAD threshold effectively over constraining the degraded audio. Further research is recommended to determine if performance (with VAD on) could be improved if settings were adjusted.

## 9.6 Discussion of Results

The research presented in this chapter has demonstrated that the influence of reverberation on ASR performance is both degrading and complex. It was shown that whilst ASR system performance declined as predicted, with greater degradation related to the size of room and complexity of reflections, the reduction in EER% was shown to be relatively insignificant for relatively low values of RT60. For example, baseline 0.005% EER to 0.02% EER (van matched SM/TA) or 0.03% EER (car unmatched SM/TA) for i-vector PLDA session 3.

### 9.6.1 Responses to Questions

**Q1 Recap:** How resilient are modern i-vector ASR systems to reverberation as opposed to the earlier GMM-UBM versions used in studies such as Castellano (1996) and Peer, Rafaely and Zigel (2008)? Further, how effective are session changes to an i-vector ASR system, based on adapting the normative data (UBM, TV, LDA+PLDA), relative to one another?

**A1:** As hypothesised, the i-vector system consistently outperformed GMM-UBM system and the EER% results are broadly in line with previous research findings e.g. Avila et al. (2015) (Table 9.5).

For the second part of Q1 and as predicted it has been shown that an increase to the size of the PLDA dataset did initiate some performance improvements for the i-vector ASR system, most specifically for baseline results. However, it was also demonstrated that these gains diminish as the PLDA size grows, assuming technical quality is consistent. Also, large RT60 values showed smaller performance gains and in certain cases a decrease with PLDA increases. This suggests that the degradation of the speech signal was of a magnitude that could not be compensated for.

**Q2 Recap: Under a given set of conditions, can we quantify the influence of reverberation on ASR performance? If so, are there any direct correlations with specific reverberation measurements such as RT60?**

**A2:** As predicted, direct mathematical correlations could not be established, due to the unknown variables inherent in the environment(s) that influence absorption and the complexity of reflections. In broad terms, environments with relatively low RT60 values and low complexity reflections/lack of hard surfaces (e.g. car and van) had very marginal influence on ASR performance. In addition, very large values of RT60 (e.g. Hall) could provide an upper RT60 threshold for which ASRs should not be deployed, since results would be deemed appreciably less reliable, especially if taken into consideration with other factors (net speech duration, band limitation, transcoding). This could suggest a potential RT60 threshold under which i-vector ASR systems are resilient enough to reverberation that they could be successfully integrated into a speaker verification workflow. In relation to the second part of the question, large RT60 values equated to poorer ASR performance, as predicted. However, the complexity of reflections and surfaces again influenced performance and acoustic assessment should be factored into the confidence of an ASR FSC task.

**Q3 Recap: Can the influence of reverberation be mitigated through:**

- **Matching conditions, i.e. RT60, for speaker model and test audio?;**
- **Adaptation or improvements to the normative data (i-vector/PLDA system) to potentially restore ASR performance?**

**A3:** As previous research suggested, matching conditions provided significant benefit over amending the normative data (assuming that the PLDA is not underspecified).

- Irrespective of the size of the room, where the speaker models and test audio are matched, it is likely that there would be less detriment to the performance of the system than where they were unmatched.
- It is likely that improving normative data relevance and adapting the feature extraction settings would partially restore performance.

## 9.6.2 Voice Activity Detection

Under conditions with very high values for RT60 (Hall and Large Hall) results showed that it is likely that the speech detection (VAD) overly prevented speech data passing to the statistical modelling phase. It was also noted that baseline EER% dropped (0.005% to 0.0017%) with the VAD set to off. Conversely, the Cllr scores were better when VAD was applied demonstrating a trade-off between overall EER% performance and accuracy. Reverberation will also likely influence delta and delta-delta measurements due to spectral smearing in the time domain. Automatically editing speech data to remove non-speech based on a single/fixed threshold could also exacerbate incorrect delta measurements if not correctly applied. To counter this, to some extent, VAD in iVocalise is integrated to maintain delta and delta-delta values through application in the feature space (i.e. MFCC). Switching off speech detection can therefore present risk in terms of EER% and under casework conditions there could be other implications. For example, if contaminant noise is present in recordings for both the test audio and an incorrect speaker model there could be a higher risk of a false verification. Full technical assessment of the audio files by an experienced analyst would be strongly recommended before removing the VAD on pre-processing (i.e. to prevent non-speech acceptance).

## 9.7 Recommendations

The following section presents several practical recommendations for consideration into the workflow integration of ASR systems.

For cars, vans or very small rooms without complex reflections and where the RT60 is low it has been demonstrated that the difference in EER%, Cllr (accuracy) and overall ASR performance degradation is relatively low in comparison with baseline performance. Therefore, if relatively small quantities of reverberation are detected during technical assessment, it might not be necessary to discard the audio, considering it below the quality threshold for ASR assessment (on this factor alone). Nevertheless, because ASR performance is degraded it supports the view that ASR analysis, particularly on reverberant speech, should be completed in conjunction and with the support of auditory phonetic comparison.

Establishing an ASR acceptability threshold for reverberant audio is difficult as it is almost impossible to objectively measure the complex influence of reverberation. The inherent variability of reverberation (proximity, surfaces and environment) also makes it difficult to provide an accurate compensation/calibration algorithm. From the experiments conducted, a strong awareness as to the influence of reverberation and the ability to make a judgement as to the depth of ASR performance degradation is recommended and should form an essential aspect of the workflow. An additional recommendation, if practicable, could take the form of impulse response measurements taken at the

test audio scene with subsequent modelling tests applied to baseline systems in order to provide objective measurements in support of outcomes (predicted EER%, Cllr, FR/FR), although this could require significant investment in terms of resources.

For moderate spaces and rooms with complex reflections it has been demonstrated that matching the test audio conditions and speaker model conditions is almost always preferable to non-matched conditions with reference to ASR performance. Again, however, this is unlikely to provide a practicable process because of the many unknown variables from the recording environment such as speaker/microphone distance, position variability, dimensions and layout of interiors and furnishings.

If recording conditions can be measured and it is practical and proportionate to do so, calculating impulse responses from the room or re-recording the speaker models in real time in the same environment over very high-quality equipment could assist with predicting the RT60 value. Theoretically this could be applied to speaker model(s) to replicate channel conditions. Whilst research has demonstrated that this is likely to provide a performance improvement, it is close to impractical with additional questions arising around validation and replication of processes. Also, the combined influence of the additional recording process on the speaker model could be disputed.

The experiments completed have shown that large reverberant spaces have a relatively strong negative influence on ASR performance. ASR results produced, under those conditions, should be treated with much caution and not considered in isolation. Other factors should also be taken into consideration too, such as the quantity of speech that passes through VAD to enrolment for both the (SM and TA) and the extent of reverberant divergence between SM and TA.

The experiments demonstrated that automatic speaker verification performance in vehicles can achieve close to baseline EER% although it must be stressed that these results are in respect solely to reverberation – and the vehicles simulated in these experiments were stationary, with the engine off. Other noise is invariably present in vehicle recordings such as engine/gearbox noise, road rumble, other traffic, CD/radio, air turbulence (windows, sunroof, air conditioning), seatbelt alarms, indicators, electrical interference or passenger babble/overlapping speech. As discussed earlier in this thesis, the proximity of the microphone, the recording equipment and any data compression applied to the recordings can also significantly degrade ASR performance. Combining reverberant and non-reverberant test audio and speaker models would not be recommended in casework conditions, as this could skew results - high false positives and low true positives (false rejects).

## 9.8 Discussion and Future Research

Returning to the spectrogram observations (Figures 9.1 to 9.4 inclusive) the question arises as to why speech smearing in the time domain, for longer reverberant conditions, has such a detrimental influence on ASR performance. It is suggested that the sub second blending of speech sounds effectively sums frequency data which affects the feature extraction for frame values (usually 10ms).

This hypothesis is supported by Shabtai, Rafaely and Zigel (2010) who suggest that when the RT60 value is greater than the short time Fourier transform that the feature vectors are smeared. They then suggest that this could cause the mean GMM values to become closer together (2010: p.41) i.e. degrading the specificity/accuracy of the statistical model.

Korany (2013) suggested that the number of coefficients could be increased in the feature extraction stage for improving performance in reverberant conditions (2013: p.6). Although this was not specifically explored in the experiments, as the Korany study was completed using a GMM-UBM system, it is suggested that it is likely that denser statistical modelling (i.e. i-vectors) would improve resilience to reverberation.

Recent research by Guzewich and Zahorian (2017) investigated the application of applying machine learning techniques to (de)reverberate material using deep neural networks (DNNs). They used an Alize (Larcher et al., 2013) i-vector ASR system and 46,200 reverberant (40 hours) and 4,620 clean speech files to test a dereverberation method (2017: p.173) based on the research from Wu et al. (2017). Guzewich and Zahorian (2017) could not replicate the results from the Wu team, which they described as 'beyond the theoretically possible' (2017: p.173). Nonetheless, Guzewich and Zahorian improved ASR performance for low T60 times (<0.20s) and recommended increasing the FFT length, which provided the greatest performance benefit (EER% not stated). They also conceded that a solution for reverberation might (at a pre-processing stage) might not benefit other speech processing (such as speech to text). It is clearly in the early stages – however, research in machine learning is likely to yield further advances.

Finally, a larger scale project to fully determine the difference between the influence of artificial reverberation and 'real world' reverberation on ASR performance would provide benefit. Whilst technically difficult, due to the large volume of recordings required, this could inform the artificial treatment of bulk speech data to bolster the UBM, TV, LDA and PLDA session, provide bespoke PLDAs or improve calibration datasets. Further advances in machine learning could also seek to treat reverberant normative sessions and speaker models as multiple object classes. The ASR could then effectively choose to make use of the closest applicable reverberant dataset(s) following machine assessment of RT60 on the incoming test audio.



# Chapter 10 Frequency Bandwidth

---

## 10.1 Introduction

This chapter examines the influence of frequency bandwidth on ASR performance. Iterative reductions in frequency bandwidth are applied to DyViS baseline data (task 1, mock police interview recordings, 44.1kHz 16bit). ASR output is then analysed and results discussed with respect to ASR performance metrics (EER% and Cllr).

The chapter begins by providing research context with an introduction to the difficulties pertaining to frequency band limited speech data and speaker recognition. A review is provided of related research literature which assisted with forming the questions and establishing experiment methodology. The research questions are then presented with associated hypotheses. A description of the experiments follows with results presented. Equal error results (EER%), log likelihood ratio (LLR) output and the cost of likelihood ratio (Cllr), or system accuracy, are examined and discussed. The research questions are revisited and responses provided.

The chapter concludes with a wider discussion offering practical recommendations for practitioners using ASR systems conducting band limited speech casework and at scale (investigative use). Proposals are also made for future research.

## 10.2 Context

This section places the experiments into the wider context of current research. The literature referenced in 10.2.2 assisted in advising the methodology and guiding the experiments conducted.

### 10.2.1 Background

The maximum frequency of standard telephony audio is constrained to a frequency bandwidth of 0-4kHz (i.e. sample rate 8kHz). It is broadly accepted that speech frequencies extend to above human hearing of approximately 16kHz to 20kHz (highest frequency) dependent on age and the individual. Although arguably the utility of high frequency speech sounds depreciates considerably towards the higher end of the frequency spectrum (>12kHz). Telephone system design significantly pre-dated computers and ASR systems, so was not engineered with machine verification in mind. As technology updates, the infrastructure and traditional means of communication adapt to alternative methods such as broadband and wi-fi. This presents opportunities to upgrade from narrowband (8kHz sample rate, 0-4kHz frequency bandwidth) to wide band (16kHz sample rate, 0-8kHz

frequency bandwidth). It is suggested that one benefit could be that the inclusion of more speech information – which is likely to improve ASR system performance.

Alternatively, it could be argued that the proximity of the microphone to speaker in telephony channel speech is predominantly good, excluding hands free or conference calls. Also, the majority of speech energy required for speaker (and indeed speech) recognition occurs well within the traditional, narrow telephony bandwidth and that increasing the frequency bandwidth further would offer only marginal ASR performance gains. In addition, almost all commercial ASR systems have optimised architecture to work predominantly in the telephony channel domain as opposed to wide band. It is assumed that this is due to market demand since the majority of ASR consumers tend to be call centres, banks and law enforcement agencies. Many of these systems already achieve relatively good EER% performance, and even better performance can be obtained for text dependent applications such as compliant speakers using telephone banking for voice authentication (e.g. a customer volunteering to repeat identical utterances) or combining speaker recognition with speech recognition.

The above argument motivated two key questions. If the frequency bandwidth is extended beyond that of telephony could any additional performance gains be exploited to better inform ASR use in casework? Conversely, when the frequency bandwidth is reduced to below telephony, how much less reliable are ASR systems with respect to performance, accuracy and precision?

## **10.2.2 Literature Review**

The effect of frequency bandwidth on speaker verification systems has been previously researched. Hayakawa and Itakura (1994) produced an early study – completing research on 5 Japanese utterances spoken by 15 males and recorded in different sessions over a year at 32kHz (sample rate). The ASR system was not specified. Their results showed that the data with the highest sample rate (i.e. 0-16kHz) provided the best recognition rates and that a ‘rich amount of speaker individual information was contained in the higher frequency band’ (1994: p.140). Hayakawa and Itakura concluded with recommending more research in this area. Misra, Ikbal and Yegnanarayana (2003) also demonstrated that EER% performance dropped when removing high frequency speech frequencies from TIMIT and NTIMIT corpus data (Jankowski et al., 1990). Their study found that 0-8kHz provided EER of 0.5% whilst 0-3.6kHz gave EER% of 6.1% (1990: p.309). A study by Gallardo, Wagner and Möller (2012) examined ASR performance over narrow band (NB) telephone channels (8kHz sample rate, 0-4kHz frequency bandwidth) and wide band (WB) (16kHz sample rate, 0-8kHz frequency bandwidth). Their research results supported previous findings by Jokic et al. (2011) and Pradhan and Prasanna (2011) that wideband speech performed significantly better in almost all experiments.

Deshpande and Holambe (2011b) examined the influence of different frequency bands by adding NOISEX-92 data, which provides real-time recordings of noise within vehicles, to the TIMIT corpus and then applying band pass filters at different intervals. They used a bespoke comparison system based on GMM classifiers (32 mixtures) and employed a new feature extraction method which was weighted towards higher frequencies (than MFCCs) called Teager Energy Operator based Cepstral Coefficients. Whilst it could be argued that this is not a direct comparative study, with MFCC based ASRs, their study demonstrated 100% identification rates on 0-8kHz and 97.33% on 0-4kHz with only 54% on 0-2kHz but 94.66% on 4-8kHz (p.195), showing much promise for alternative feature extraction methods.

The Pradhan and Prasanna (2011) study further demonstrated that the performance improvement was greater for females than males (see Table 10.1).

**Table 10.1:** Results from NB and WB ASR performance. Pradhan and Prasanna (2011)

Narrow Band EER%	Wide Band EER%
Male baseline: <b>9.49</b>	Male: <b>7.34</b>
Female baseline: <b>10.52</b>	Female: <b>4.00</b>

Pradhan and Prasanna (2011) proposed that the reason for the gender performance differential was likely due to the higher pitch and formants of female speech than males. This is a logical and plausible hypothesis. In addition, for the third key component (universal background model), their normative data was carefully selected to ensure robust gender balance (17 male speakers and 17 female speakers with five hours speech from each group). Pradhan and Prasanna (2011) also provided evidence that the performance improvement for wideband in comparison to narrowband speech held true in almost all instances, including relatively mismatched and/or noisy conditions.

Gallardo, Wagner and Möller (2012) examined 51,200 cross comparisons from the ANDOSL and AusTalk databases, running the experiment multiple times through a bespoke ASR system to determine EER% output under 5 different conditions. All audio files in the experiments were passed through the same processes to avoid channel mismatch. The ASR system used was a Matlab 7.13 R2011b and OS code system, rather than a commercial off the shelf (COTS) product. The group utilised open source code to perform a standard MFCC extraction with a GMM classifier. For the purposes of their research, WB was categorised as 50Hz to 7kHz and NB as 300Hz to 3.4kHz. A bespoke UBM was necessary and so created from speaker data representing the conditions of the test. It is suggested that this is very likely to have artificially raised the performance of the system, due to the specificity of the normative speech data. So, whilst the methodology was clearly applied under a research context, performance results would not transition to casework ASR analysis – where the normative data are not specifically tailored to the casework conditions. Nevertheless, their results

demonstrated that the performance and accuracy of verification systems did increase when wideband (WB) speech signals were used over narrow band (NB) with 12.03 EER% for (Adaptive Multi Rate) AMR-WB compared to 18.53 EER% for AMR-NB demonstrating a 64.92% improvement. However, it could be argued that the Gallardo group's results were also affected by additional variables such as transcoding (data compression). For example, the AMR-WB had an effective bit rate of 23.05kbps whilst the AMR-NB was at 4.75kbps. For secondary trials, the G.722WB (12.45% EER) and G.711NB (16.45% EER) tests conducted both used 64kbps bit rate, though it is possible that the performance differential could also be partially influenced by other codec differences between G.711 and G.722.

Besacier and Bonastre (2000) demonstrated that, for 630 speakers, the most important frequencies for speaker verification systems were not evenly distributed. Low frequency bands under 600Hz and high frequency bands over 2kHz were found to be more speaker specific than those in the middle range. This was supported by research from Orman and Aslan (2001) examining 16kHz speech for 462 speakers. They also showed that certain frequency bands were more pertinent to automatic speaker verification systems than others, suggesting that key frequency ranges were from 0Hz to 1kHz and 3kHz–4.5kHz acknowledging that the lower frequencies of speech do not descend to 0Hz (approximately >80Hz dependent on gender, age, language, health etc.). Whilst it should be noted that the ASR systems in both these studies used GMM-UBM architecture rather than a more modern i-vector approach, this research was of particular interest with respect to the extension of frequency bandwidth beyond standard telephony.

In reference to the research literature several key technical points were extrapolated which influenced the research questions and experiment methodology in this chapter.

### **Transcoding**

As transcoding can influence frequency bandwidth, codec type and settings should be considered and preserved with respect to the original recording(s). In frequency bandwidth experiments it is important not to additionally transcode so as to avoid conflating variables. By extension, in casework for example, if transcoding is mandatory (e.g. audio submitted in a codec that is incompatible with the ASR) then the transcoding process should be factored into analysis of ASR results and documented. More broadly, if transcoding is applied it should ideally be without any frequency bandwidth limitation and with zero data compression (i.e. lossless).

### **Population Data**

The UBM proposed for use in this chapter was re-examined and regarded as unsuitable with regard to frequency bandwidth (i.e. speech frequencies not present from 4kHz to 12kHz). A suitable UBM was required to reflect the wide band interview channel. To eliminate potential contamination of

results, it was determined that no part of the DyViS corpus should populate any part of the normative set.

### **Multiple Bandwidth Tests**

From previous research conducted it was determined that a greater number of experiments at multiple frequency bandwidth settings would improve the detail of results. This might then assist in terms of observing smaller changes, should they occur. For example, iterative steps of frequency range limitation could inform a series of LR and/or zoo plots, which might show the rate of performance change through inter-speaker distance movement. It was also noted that when limiting channel bandwidth with a low pass filter (LPF) it should be applied in a way that does not simultaneously apply a high pass filter (HPF).

### **Quality Control**

The experiments required the generation of hundreds of thousands of treated files. Whilst it was impractical to check every single file, the technical quality of recordings was carefully spot checked (approximately 5-10%) to check for unwanted aliasing and/or artefacts or inconsistencies in the batch process.

### **Net Duration**

It was noted that the speech data samples used to populate the normative data (UBM) in the Gallardo, Wagner and Möller (2012) research were relatively brief (5 seconds) and this was under the OWR and ENFSI recommended sample times for ASR analysis (approximately 20s for each of the SM and TA audio files). It was determined that the use of longer speech samples (1m speaker model and multiple 1m test audio files) would decrease the potential influence of net duration on the experiment and mitigate against conflating variables.

### **System Architecture**

From the research literature, it was hypothesised that i-vector systems were likely to be more robust to channel bandwidth degradation than GMM-UBM ASRs. An experiment to test the two ASR systems should be conducted to examine and quantify this.

### **Very Low Bandwidth Speech**

Little research was found on sample rate/frequency bandwidth below standard telephony channels and ASR performance. An experiment could inform the extent of performance deterioration below the upper frequency limit of 4kHz. To place an experiment into a practical context, a relevant casework example would be ASR speaker comparison conducted on speech data from push to talk radio (PTTR) systems (or walkie-talkies), which generally constrain frequency below an upper limit of 3.5kHz.

## 10.3 Questions and Hypotheses

The experiments in this chapter were generated to address four key questions.

**Q1 Does ASR performance noticeably improve relative to baseline when the frequency bandwidth is extended beyond telephony? If so, what is the optimum frequency bandwidth for ASR performance?**

**H1** Wider band recordings (0Hz to 8kHz) should provide ASR performance improvements over constrained telephony recordings (0Hz to 4kHz) due to the greater quantity of speech data captured for statistical modelling. However, as the majority of speech energy exists within the telephony recording range the performance increase is likely to be marginal. It is suggested that neither GMM-UBM nor i-vector/UBM, TV, LDA and PLDA systems will be optimised easily to work on wideband speech data by default – and will require adaptation. This is because most ASR systems are optimised to work on narrow band telephony data with respect to feature extraction method and normative data composition. In addition, speech energy diminishes in dynamic range for higher frequencies (8kHz to approximately 12kHz). Also, the experiments conducted include data solely from male speakers (i.e. lower average frequency range) so it is suggested that any improvement in EER% is likely to plateau rather than continuing to improve – and results will not extrapolate directly to female speech.

**Q2 Does an i-vector/UBM, TV, LDA and PLDA ASR system offer significant performance advantages over a GMM-UBM system when the frequency bandwidth is extended?**

**H2** Broadly speaking, i-vector/UBM, TV, LDA and PLDA ASR systems outperform GMM-UBM systems. Whilst the MFCC extraction process remains similar, the improvements in the statistical modelling process should positively influence performance for i-vector systems.

**Q3 Many ASR systems automatically downsample audio files as they are imported, to a frequency bandwidth 0-4kHz (sample rate 8kHz). OWR Vocalise and iVocalise ASR software systems provide the operator with the opportunity to adjust the frequency bandwidth (minimum and maximum settings) for the MFCC feature extraction stage and allow the configuration of normative data. Can performance advantages therefore be found in terms of matching frequency bandwidth for speaker models and test audio?**

- **If we applied the same channel bandwidth limitation to both the questioned audio and speaker model, how would ASR performance vary against baseline?**
- **If iterative bandwidth degradation was applied to the test audio but wide band speaker models were used, how would ASR performance vary against baseline?**

**H3** Matching SM and TA has been shown to predominantly improve ASR performance, so frequency band limitation applied to both the test audio and speaker models should provide better ASR performance as the data is effectively complete on both sides of any comparison. Conversely,

poorer performance should occur where there is variation between TA and SM with respect to frequency bandwidth – and this is likely to degrade further on higher divergence between SM and TA.

**Q4** If the frequency bandwidth is significantly reduced below that of standard telephony what implications would that have for ASR performance?

**H4** ASR Performance degradation is likely to occur as speech energy is removed from an area of the spectrum shown to have speaker-discriminating potential. This is in respect of both consonants and vocalic segments. The vocalic information lost, below the telephony bandwidth (0 to 4kHz, 8kHz SR) would include F4, and, as frequency bandwidth is increasingly constrained, F3. This would be consistent with research completed by Gold, French and Harrison (2013). Potentially important consonantal information to be lost includes the energy loci and distributions occurring with anterior fricative consonants (Kavanagh, 2012). While the MFCC feature extraction process is insensitive to individual segmental features, the compound effects of removing speaker discriminatory energy patterns are likely to result in confusion of speakers and diminished ASR system performance. In addition, unmatched conditions (between SM and TA) would likely increase confusion, with less speech energy present in one than the other. Understanding any performance tipping points, in terms of higher frequency cut-offs, might assist in informing thresholds for when ASR use would not be recommended.

## **10.4 Methodology**

The method as outlined in chapter 5 was observed with the following changes.

### **10.4.1 Baseline Corpus**

Speech data from DyViS task 1 (mock interview, 44.1kHz 16bit, 100 speakers) was used to generate both speaker models and test audio files. The audio files were edited as follows. Speaker models were created using 1 minute of speech data with the remaining net speech divided into 3 further extracts for testing (i.e. 1m, 1m, residual). When analysed using an ASR system this then produced 30,000 cross comparisons with 300 same speaker scores and 27,000 different speaker comparisons.

### **10.4.2 Automatic Speaker Recognition Systems and Additional Materials**

It was determined that the frequency band limitation should be completed in controlled, iterative steps. This was to analyse performance results in detail with regard to potential zoo plot movement, to seek possible cliff edge effects (Q4) and to identify optimum performance conditions (Q1).

Batch processing was required due to the quantity of files. Several software solutions were identified. These were assessed as to suitability, practicality and output quality. Tests were then conducted with particular focus on the quality of sample rate conversion. Output was dip sampled (approximately 10%) for aliasing or any other acoustic artefacts using spectrogram analysis. Several applications were ruled out due to the potential for acoustic contamination. Others were rejected due to poor workflow (e.g. number of steps required and/or processing speed).

The iZotope RX6 Advanced application ([Izotope.com](http://izotope.com)) was shown to have an extremely high quality SRC output, without introducing unwanted artefacts. The application also utilises a brick wall, high pass filter and was found to be extremely fast when batch processing. From trials, the steepness of the high pass filter was essential in ensuring that frequencies close to the HPF cut off point were not affected by gain reduction or aliasing. Any introduction of a slope at the cut-off point would diminish speech frequencies rather than eliminating them.

An additional requirement was that batches should be converted incrementally in iterative decreasing steps. To mitigate for data contamination from cumulative conversions the process was not applied in succession. Each time band limitation was applied it was to the first generation audio data rather than to that produced by the previous step.

### **10.4.3 Test Audio and Speaker Models**

Two batches of test data (speaker models and test audio files) were created from the DyViS task 1 (mock interview) data. In the first set, both the speaker models and test audio were treated simultaneously with respect to frequency bandwidth limitation. The highest frequency bandwidth setting was 0-16kHz (i.e. sample rate of 32kHz). This setting was deliberately chosen to exceed the frequency range of speech to determine if any non-speech, high frequencies captured might negatively affect ASR performance (e.g. neon light hum). The lowest frequency bandwidth was set to 0-2.5kHz (i.e. sample rate of 5kHz). This was chosen to simulate the type of channel occurring on a typical push to talk radio (PTTR) system.

In the first set of data, frequency intervals were selected at 1kHz creating x11 incremental steps between the highest and lowest sample rates for both the speaker models and test audio (i.e. matched conditions). In the second set, the speaker models were consistently held at the highest frequency bandwidth/sample rate and the test audio was degraded in x11, 1kHz incremental steps (i.e. unmatched conditions).



#### 10.4.4 Normative Data

In conjunction with the speaker model (SM) and test audio (TA) the third data set, often unseen to the operator, is the normative data. The normative data informs the ASR as to what speech is and provides statistically mean values for all features extracted from the population used to compile it.

From completing background research and compiling the research questions, the OWR ASR systems were further examined in respect to normative data. It was noted that all of the underlying normative data, within the default ASR configurations for Vocalise, was optimised for telephony bandwidth (0-4kHz). It was determined that the lack of upper frequency speech data, in the normative set, could potentially influence experiment results and negate any benefit of extending channel bandwidth. Put simply, the ASR would have no data points beyond 4kHz to inform the statistical model as to reference values. The UBM created for use in other chapters was also deemed unsuitable for the frequency bandwidth experiments due to (GSM) transcoding and microphone proximity. New reference data was therefore constructed for both GMM-UBM and i-vector/PLDA versions.

For the Vocalise ASR system, the GMM-UBM was created using in domain channel audio data from a similar speaker demographic as DyViS (i.e. interview, male, SSBE, 18-25 years). Files were carefully checked acoustically, using spectrograms. Criteria assessed included poor signal to noise ratio, mains hum and other acoustic artefacts which could adversely influence results and many speech files were rejected. Eighty-nine speakers, with interview speech session files, were chosen from the SPOKE database to use as normative data (GMM-UBM). The feature extraction settings on Vocalise were then adjusted to extend the speech frequency limits, enabling wide band comparison.

As discussed in previous chapters, the normative data for the i-vector version takes the form of a multiple stepped process and it is possible to train using different data for each component (UBM, TV and LDA+PLDA). As the population dataset used to train the model was extremely large – multiple training sets for each of the models (UBM, TV and LDA+ PLDA) was not required and one set was used for all components. This was consistent with advice, from OWR, stating the overall performance of the system benefits of using the same data to train the UBM, TV and LDA+PLDA.

#### 10.4.5 Automatic Speaker Recognition Systems

See Appendix G for additional details. The two systems compared were:

- i. OWR Vocalise 1, GMM-UBM system: version 1.5.0.1190 (bespoke UBM)
- ii. OWR iVocalise, i-vector/UBM, TV, LDA+PLDA system: version 2.1.0.1366  
PLDA set '*2016A-1024-D-CMS-Large-VAD-NoDyViS-20Apr16*'

The TV (total variability) was set to 400 dimensions, the PLDA set to 200 dimensions and 10 train cycles.

Further details are provided in Appendix G. Note that the ASR systems were adapted with bespoke normative sets, neither containing DyViS material.

Bio-Metrics version 1.8.0.704 was used for generating performance data graphs and charts from the .csv output files (EER%, H0, H1, Cllr and for graphing and plotting results).

### 10.4.6 Data and List of Experiments

DyViS corpus, task 1, studio quality, mock Police interviews, 100 speakers. Edited to produce 100 speaker model files (SM) and 3 x 100 test audio (TA) files (29,700 TN and 300 TP). Fifteen frequency band limited comparison sets created in iterative steps. From 0-3kHz (SR 06kHz) to 0-16kHz (SR 32kHz) inclusive.

Experiments were set as follows:

- i. **Batch test 1** iVocalise i-vector/PLDA. Matched conditions.  
Speaker models and test audio have same settings, e.g. SR06kHz SM to SR06kHz TA.
- ii. **Batch test 2** iVocalise i-vector/UBM, TV, LDA+PLDA. Unmatched conditions.  
Speaker models fixed WB with variable test audio, e.g. SR32kHz SM to SR06kHz TA.
- iii. **Batch test 3** Vocalise GMM-UBM. Matched conditions.  
Speaker models match test audio band settings
- iv. **Batch test 4** Vocalise GMM-UBM. Unmatched conditions.

## 10.5 Results

See section 9.5 for additional explanation of H0 mean, H1 mean, H0 SD, H1 SD, FAR and FRR as referenced in the results Tables presented.

## 10.5.1 iVocalise, i-vector System Results

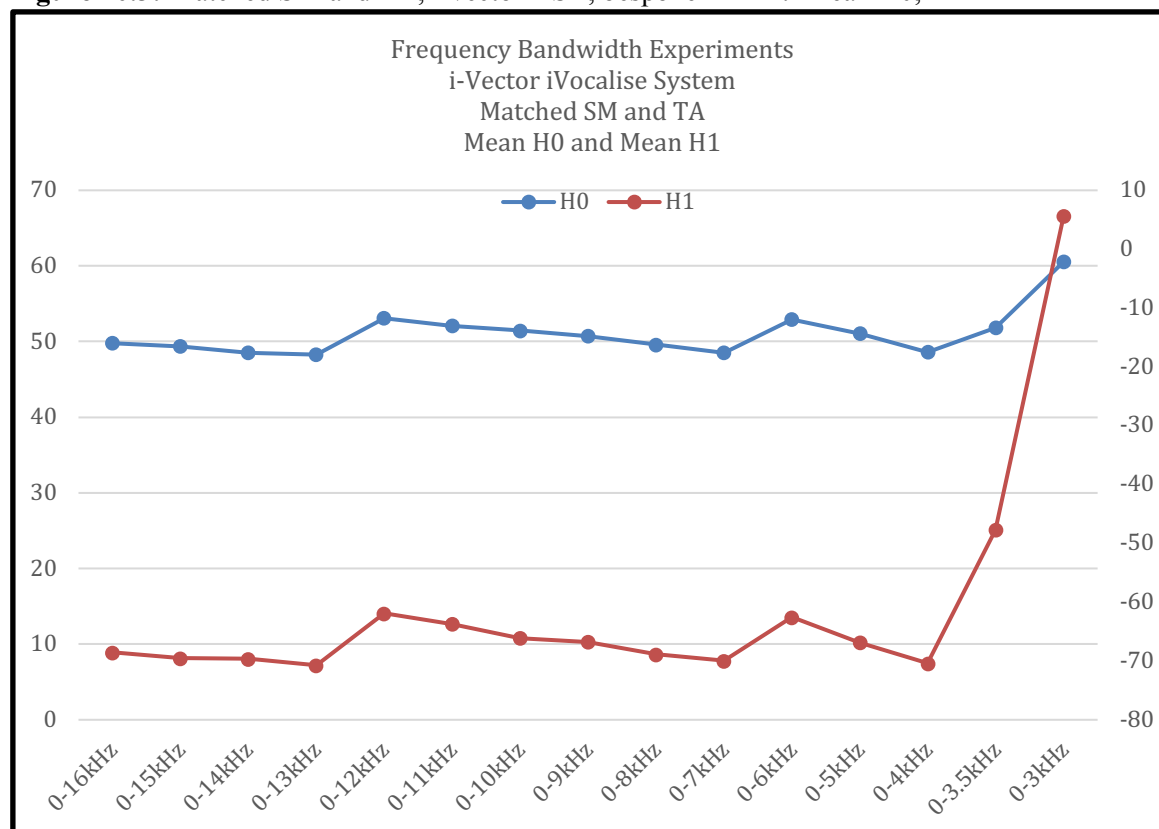
**Table 10.2:** Matched SM and TA. iVocalise ASR, bespoke PLDA Results

Frequency Bandwidth h	EER%	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR 100 %	FAR, FRR 1,000 %	FAR, FRR 10,000 %
0-16kHz	0.0320	0.01	49.79	-68.60	12.59	27.80	0.00	0.00	0.67
0-15kHz	0.0556	0.01	49.36	-69.57	12.74	27.75	0.00	0.33	1.34
0-14kHz	0.0505	0.01	48.52	-69.68	12.83	27.84	0.00	0.10	3.67
0-13kHz	0.0741	0.01	48.27	-70.77	12.77	27.87	0.00	0.33	1.68
0-12kHz	0.0404	0.02	53.10	-61.95	12.32	27.24	0.00	0.00	1.33
<b>0-11kHz</b>	<b>0.0269</b>	0.02	52.09	-63.74	12.18	27.57	0.00	0.00	1.33
0-10kHz	0.0286	0.02	51.47	-66.13	12.32	27.69	0.00	0.00	2.33
0-9kHz	0.2559	0.02	50.75	-66.79	12.69	27.85	0.00	0.33	2.00
0-8kHz	0.0404	0.01	49.58	-68.91	12.91	27.82	0.00	0.00	2.33
0-7kHz	0.0707	0.01	48.55	-70.03	12.93	27.91	0.00	0.33	3.00
0-6kHz	0.0320	0.02	52.93	-62.63	12.06	27.35	0.00	0.00	1.00
0-5kHz	0.0320	0.01	51.05	-66.93	12.64	27.70	0.00	0.00	2.01
0-4kHz	0.0421	0.01	48.63	-70.43	12.86	27.96	0.00	0.00	1.33
0-3.5kHz	0.3300	0.12	51.81	-47.77	12.02	27.26	0.00	2.10	6.33
0-3kHz	1.0067	7.25	60.55	5.64	8.22	16.74	1.00	5.72	23.67

**Optimum EER% performance in bold**

Note the poor Cllr performance (accuracy) for the 0-3kHz test, despite relatively good EER%

**Figure 10.3:** Matched SM and TA, i-vector ASR, bespoke PLDA. Mean H0, H1



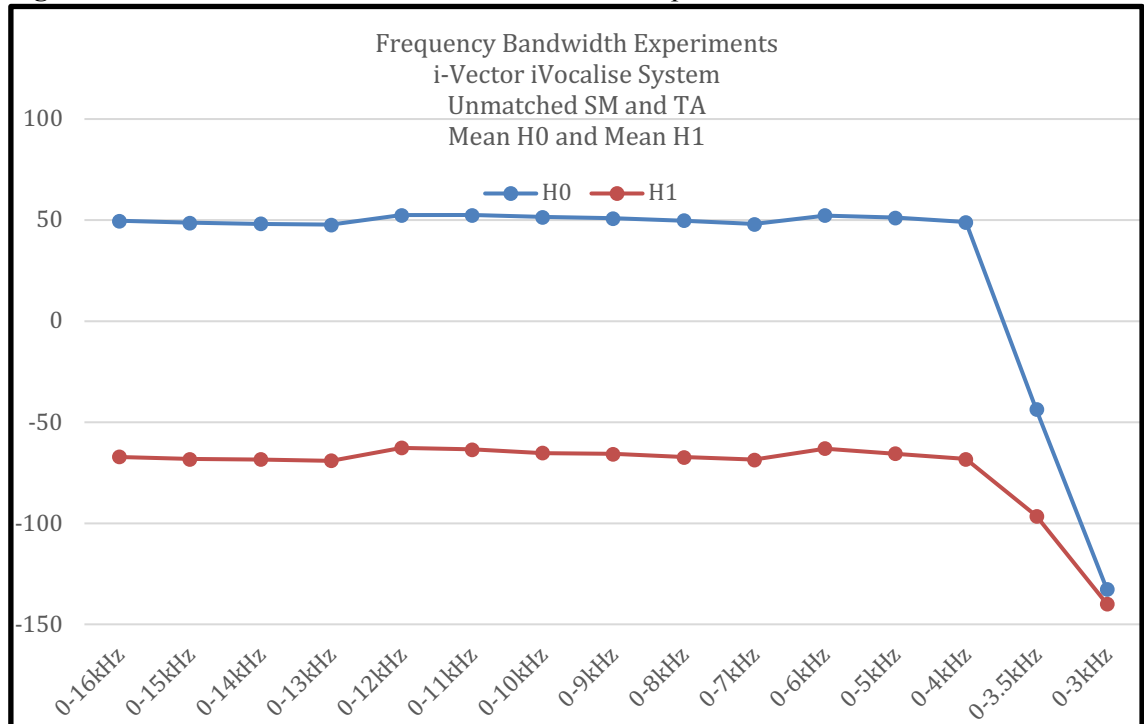
**Table 10.4:** Unmatched SM and TA, i-vector ASR, bespoke PLDA SM fixed at 44kHz. Variable TA.

Frequency Bandwidth	EER%	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
0-16kHz	<b>0.0438</b>	0.01	49.67	-67.03	12.63	27.71	0.00	0.00	2.67
0-15kHz	<b>0.2525</b>	0.01	48.68	-68.20	12.74	27.83	0.00	0.33	2.00
0-14kHz	<b>0.0707</b>	0.01	48.18	-68.37	12.74	27.83	0.00	0.33	3.00
0-13kHz	<b>0.2593</b>	0.01	47.78	-69.02	12.72	27.85	0.00	0.33	3.00
0-12kHz	<b>0.0522</b>	0.02	52.52	-62.62	12.32	27.36	0.00	0.33	1.33
0-11kHz	<b>0.0455</b>	0.02	52.40	-63.43	12.28	27.52	0.00	0.00	1.33
0-10kHz	<b>0.0606</b>	0.02	51.48	-65.09	12.48	27.67	0.00	0.33	2.34
0-9kHz	<b>0.0673</b>	0.02	50.82	-65.62	12.54	27.70	0.00	0.67	2.00
0-8kHz	<b>0.0572</b>	0.01	49.72	-67.19	12.67	27.77	0.00	0.33	2.01
0-7kHz	<b>0.0724</b>	0.01	48.04	-68.53	12.86	27.76	0.00	0.67	3.00
0-6kHz	<b>0.0455</b>	0.02	52.36	-62.94	12.27	27.35	0.00	0.00	1.67
0-5kHz	<b>0.0606</b>	0.02	51.25	-65.51	12.60	27.64	0.00	0.67	1.68
<b>0-4kHz</b>	<b>0.0370</b>	0.01	49.03	-68.23	12.78	27.75	0.00	0.00	1.84
0-3.5kHz	16.302	31.46	-43.56	-96.37	20.49	31.35	79.75	97.0	99.67
0-3kHz	43.803	95.63	-132.57	-139.77	24.00	25.54	99.67	100	100

**Optimum EER% performance in bold**

Note poorer EER% and Cllr performance in comparison to matched conditions.

**Figure 10.5:** Unmatched SM and TA, i-vector ASR, bespoke PLDA H0, H1 SD



## 10.5.2 Vocalise, Gaussian Mixture Model System

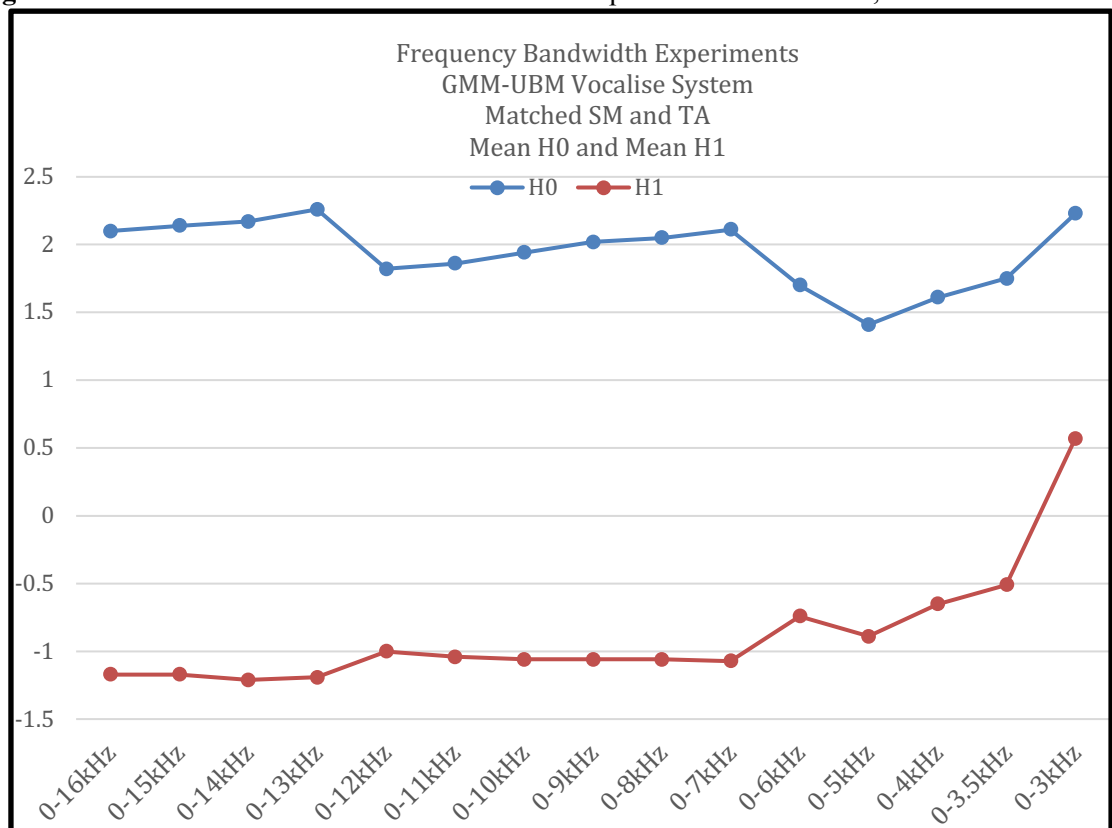
### Results

Table 10.6: Matched SM and TA. Vocalise GMM-UBM, bespoke UBM

Frequency Bandwidth	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
0-16kHz	0.32	0.83	2.10	-1.17	0.52	0.83	0.00	1.43	8.35
0-15kHz	0.33	0.84	2.14	-1.17	0.53	0.84	0.00	2.00	10.33
<b>0-14kHz</b>	<b>0.27</b>	0.86	2.17	-1.21	0.53	0.86	0.00	1.43	7.67
0-13kHz	0.31	0.87	2.26	-1.19	0.54	0.87	0.00	1.00	9.34
0-12kHz	0.33	0.73	1.82	-1.00	0.47	0.73	0.00	2.10	11.68
0-11kHz	0.37	0.76	1.86	-1.04	0.49	0.76	0.00	2.00	12.36
0-10kHz	0.61	0.77	1.94	-1.06	0.50	0.77	0.00	2.67	8.35
0-9kHz	0.67	0.79	2.02	-1.06	0.51	0.79	0.00	2.67	16.01
0-8kHz	0.58	0.35	2.05	-1.06	0.49	0.80	0.00	3.11	9.02
0-7kHz	0.65	0.82	2.11	-1.07	0.52	0.81	0.00	3.43	12.00
0-6kHz	0.68	0.44	1.70	-0.74	0.49	0.63	0.67	12.61	46.69
0-5kHz	2.66	0.45	1.41	-0.89	0.60	0.63	6.00	28.10	84.67
0-4kHz	2.43	0.48	1.61	-0.65	0.60	0.62	4.18	22.78	61.01
0-3.5kHz	2.67	0.50	1.75	-0.51	0.62	0.64	6.00	38.67	71.33
0-3kHz	5.69	0.84	2.23	0.57	0.59	0.53	29.00	67.33	80.67

Optimum EER% performance in bold

Figure 10.7: Matched SM and TA. GMM-UBM bespoke UBM. Mean H0, H1

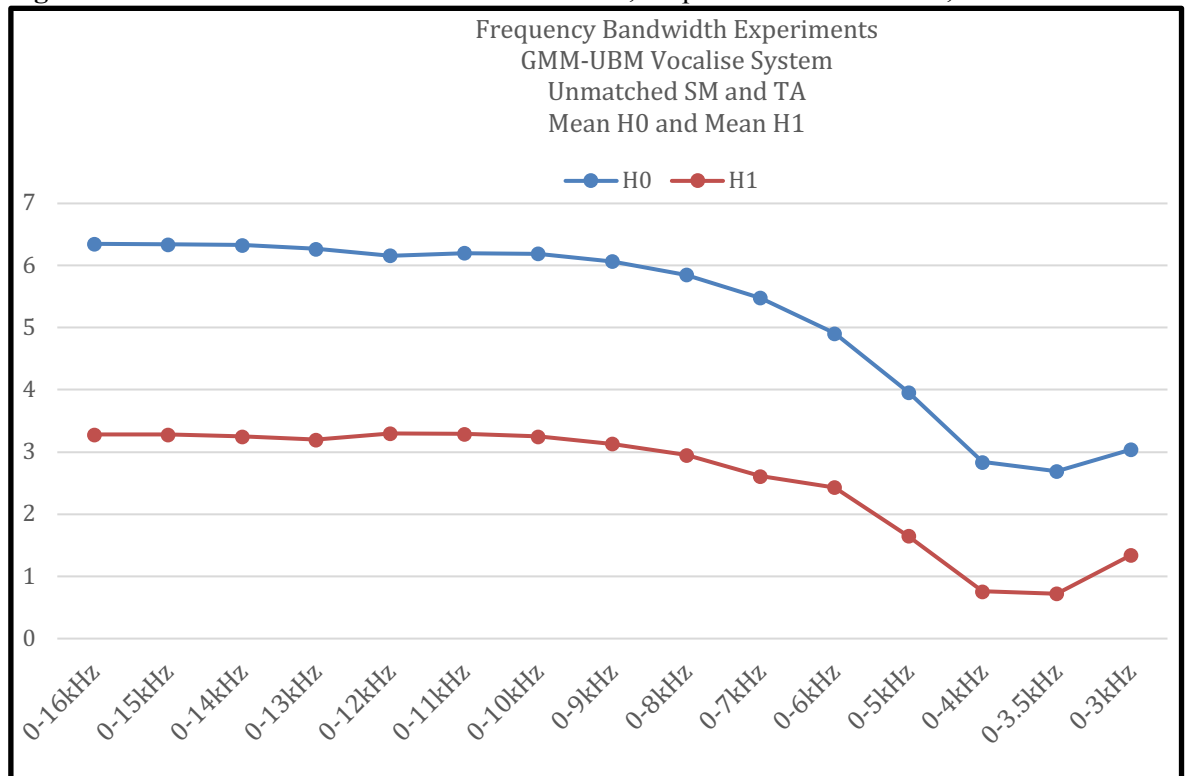


**Table 10.8:** Unmatched SM and TA. GMM-UBM, bespoke UBM.

Frequency Bandwidth	EER %	Cllr	H0 Mean	H1 Mean	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
0-16kHz	5.66	2.41	6.35	3.28	0.89	1.01	19.67	45.63	76.02
0-15kHz	5.93	2.41	6.34	3.28	0.90	1.02	20.76	45.05	76.33
0-14kHz	6.16	2.39	6.33	3.25	0.91	1.05	21.00	46.66	70.69
0-13kHz	6.28	2.36	6.27	3.20	0.92	1.05	20.33	45.48	74.01
0-12kHz	7.69	2.42	6.16	3.30	0.95	1.02	26.00	53.88	74.68
0-11kHz	7.76	2.42	6.20	3.29	0.97	1.05	29.53	56.43	78.01
0-10kHz	7.31	2.39	6.19	3.25	0.98	1.05	31.15	57.22	73.00
0-9kHz	7.34	2.30	6.07	3.13	0.97	1.05	31.70	57.67	73.02
0-8kHz	7.10	2.19	5.85	2.95	0.96	1.03	31.67	60.00	78.33
0-7kHz	6.68	1.97	5.48	2.61	0.93	1.01	28.74	55.43	76.03
0-6kHz	8.00	1.85	4.91	2.43	0.89	1.85	32.12	52.55	78.17
0-5kHz	6.24	1.36	3.96	1.65	0.78	0.74	29.26	49.77	64.70
<b>0-4kHz</b>	<b>2.18</b>	0.90	2.84	0.76	0.54	0.56	5.67	16.99	40.01
0-3.5kHz	2.64	0.88	2.69	0.72	0.50	0.56	6.25	19.43	41.52
0-3kHz	5.35	1.19	3.04	1.34	0.55	0.54	18.67	44.65	66.35

Optimum EER% performance in bold

**Figure 10.8b:** Unmatched SM and TA. GMM-UBM, bespoke UBM. Mean H0, H1



### 10.5.3 Zoo Plots

The zoo plots and a set of LR plots from the GMM-UBM system were also placed into two.gif animation files, submitted in support of this thesis:

- i. NASH\_108045162\_GMM-UBMAnimation\_FreqBandwidth\_matched\_LRPlots.gif
- ii. NASH\_108045162\_GMM-UBMAnimation\_FreqBandwidth\_matched\_Zoos.gif

The zoo plot animation (fixed axis values) demonstrates the performance degradation as frequency bandwidth is constrained – with a noticeable shift of speaker points to the lower quadrants (poorer performance) at the lowest settings.

The LR plot animation (non-fixed axis) demonstrates overall steps in ASR performance, where true positive and negative scores degrade with frequency bandwidth, marginally improve and then degrade again. The reason for this is not known - one explanation could be that the optimum positioning of the MFCC filters is shifting against (fixed) formant values as frequency bandwidth is constrained and that there are also certain ‘sweet spots’ re the relevance of the normative data (i.e. 8kHz sample rate files) however, further research is required.

Five zoo plots are presented (see also Appendix E).

Figure 10.9: Zoo plot re frequency bandwidth, 0-16kHz, SR32kHz Matched SM and TA.

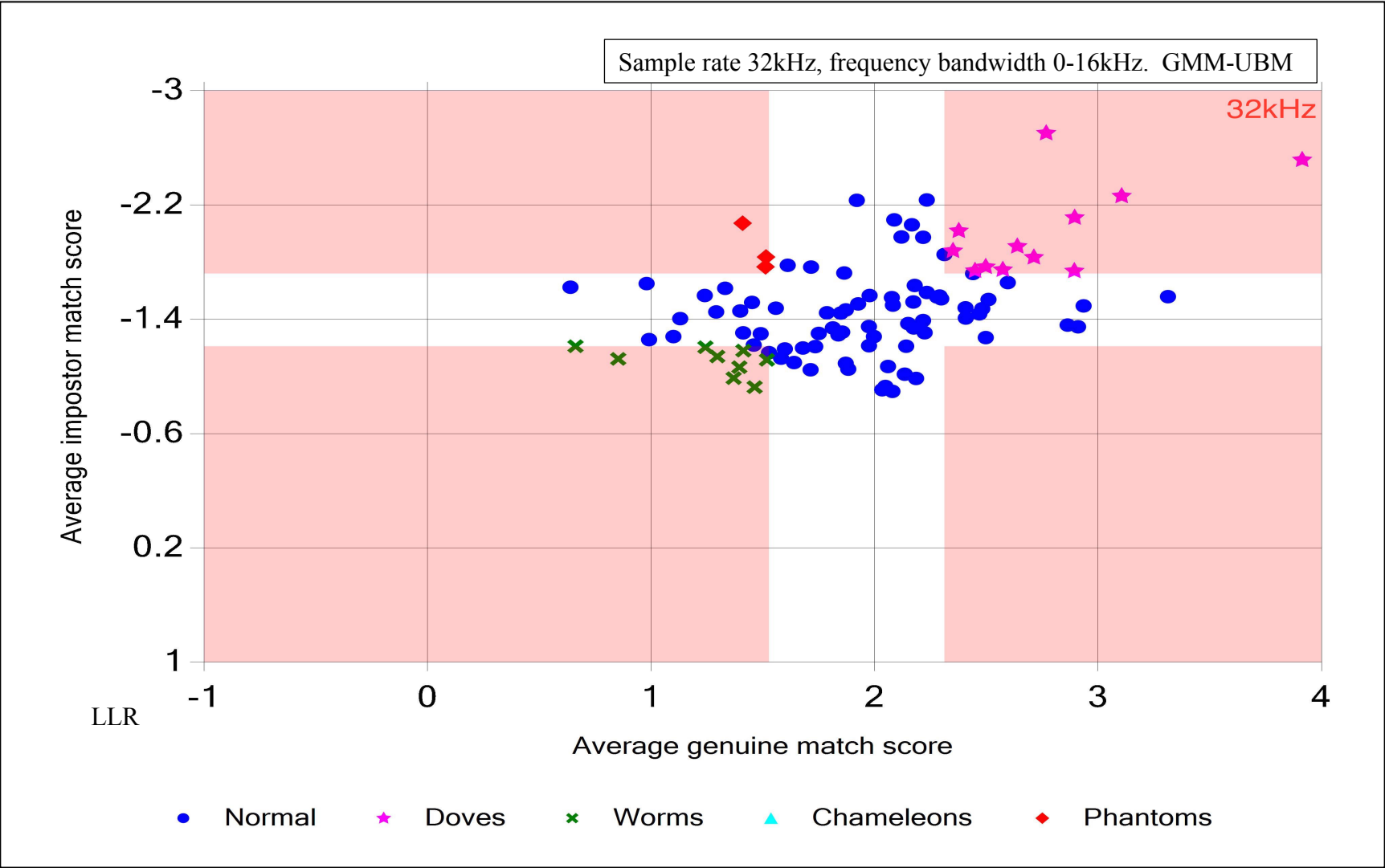
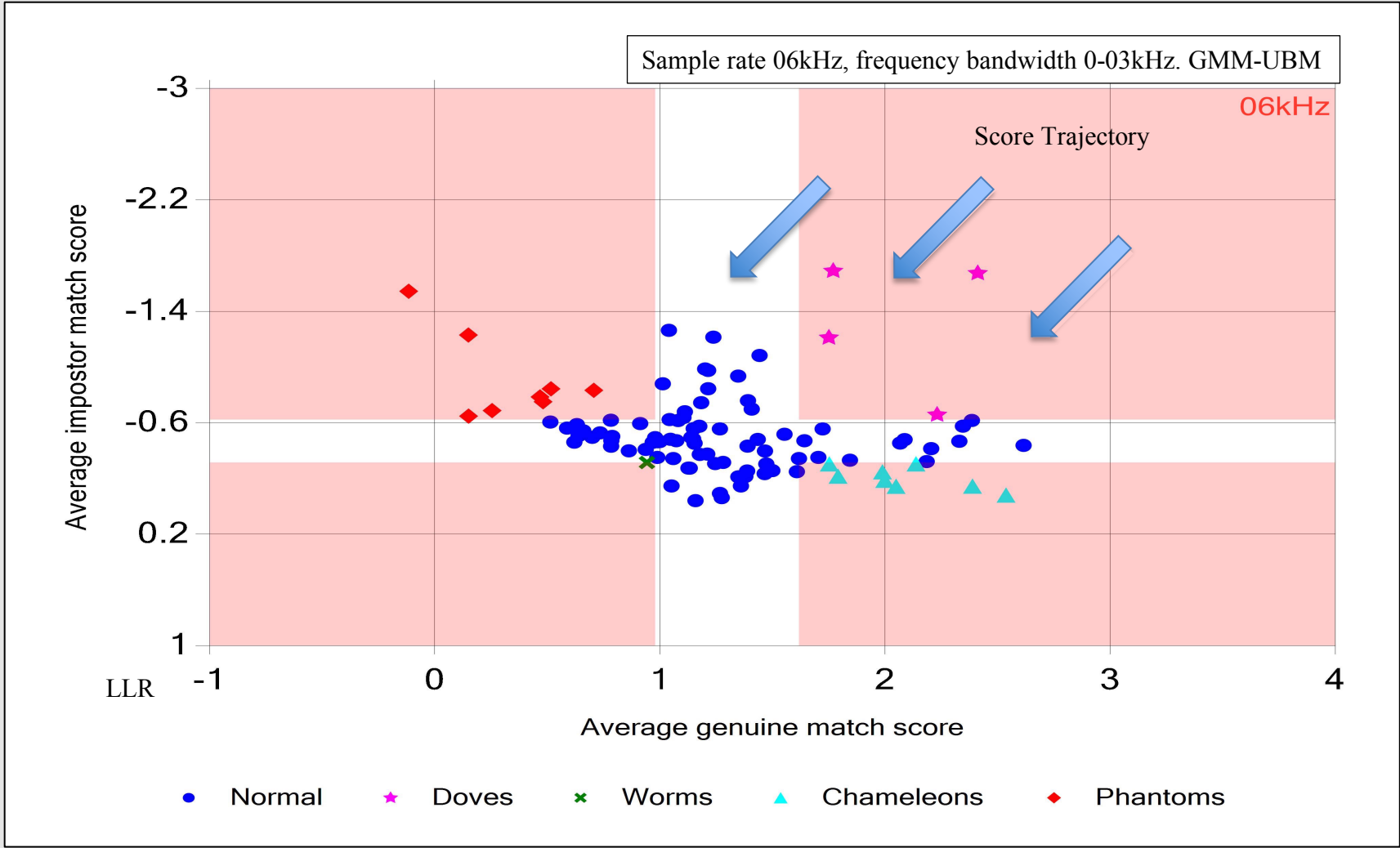
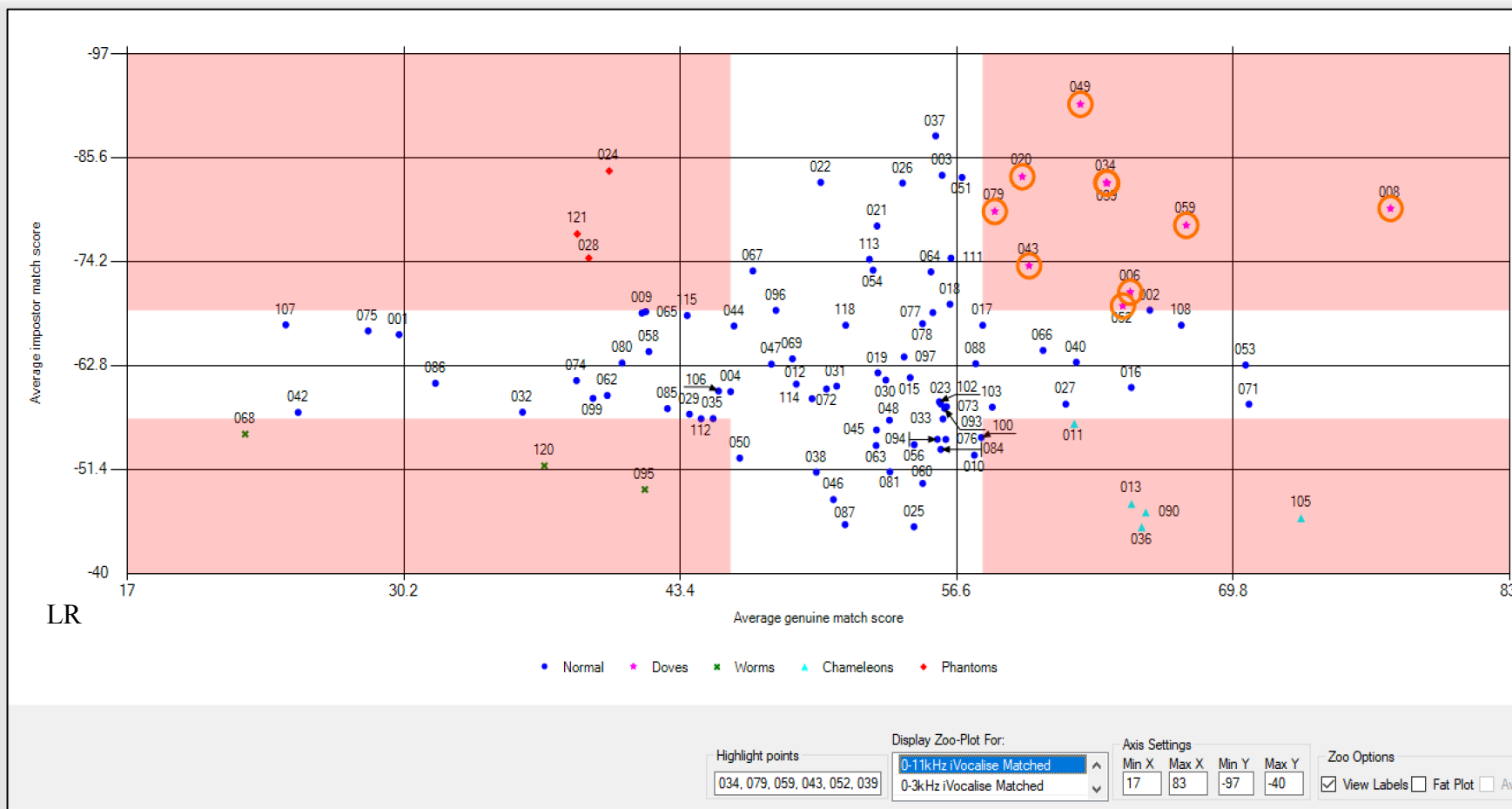




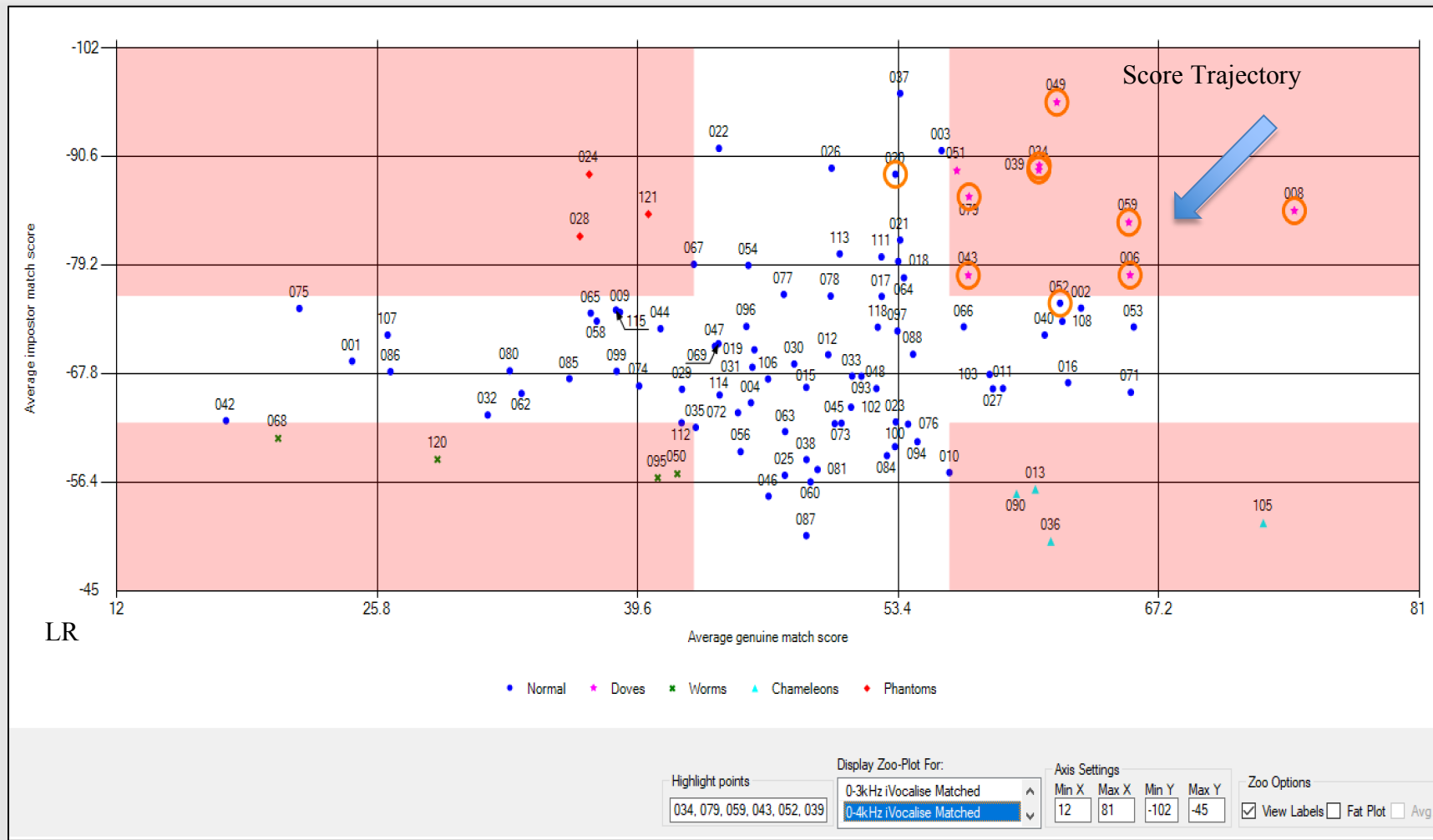
Figure 10.10: Zoo plot re frequency bandwidth, 0-03kHz, SR06kHz Matched SM and TA.



**Figure 10.11:** Zoo plot iVocalise 0-11kHz Matched SM and TA. Dove speakers highlighted.

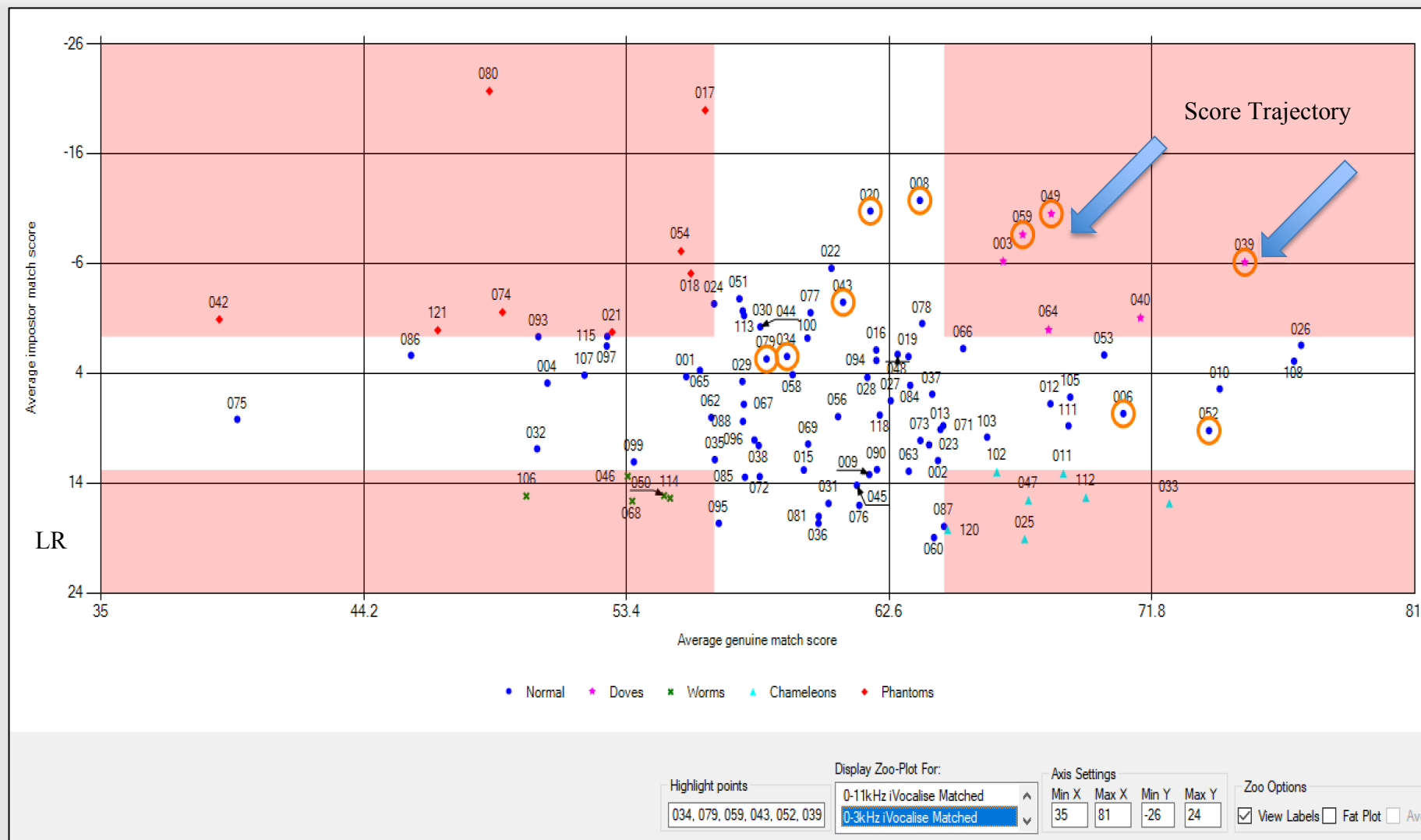


**Figure 10.12:** Zoo plot iVocalise 0-4kHz Matched. Dove speakers from 0-11kHz test



Note axis change in comparison to Figure 10.11

**Figure 10.13:** Zoo plot iVocalise 0-3kHz Matched. Dove speakers from 0-11kHz test



Match scores decreased and imposter scores increased as frequency bandwidth was constrained on both ASR systems (Figures 10.9, 10.10 and Appendix E). This was noticeable using Zoo plot visualization with data points gradually moving in iterative steps, from the upper left quadrant to the lower right quadrant as LLRs lowered overall and score separation deteriorated. Also visible (Figures 10.9 and 10.10) was the increase in the number of speakers that were difficult for the ASR to verify - Chameleons (pale Blue) elevated from 0 to 8 and Phantoms (Red) from 3 to 8.

There was a fall in the number of speakers easily recognised by the system with Doves (Pink) reduced from 12 at the highest frequency bandwidth (0-16kHz, 30kHzSR) to 4 at the lowest (0-3kHz, 6kHz SR) on the iVocalise (matched). Note also the similarity in positioning for the 0-4kHz system and the difference in final Dove position at the lowest frequency bandwidth (iVocalise frequency bandwidth 0-3kHz, SR 06kHz, matched). Figures 10.11, 12 and 13 further illustrate the Dove speaker positions, i.e. the best performing speakers on the optimal ASR system (iVocalise 0-11kHz frequency bandwidth, SR 22kHz, matched). Dove positioning (speakers 008, 006, 052, 034, 039, 049, 020, 079, 043 and 059) shifted towards the lower left quartile overall as the frequency bandwidth was constrained with speakers 006, 034, 043, 052 and 079 degrading into the central/normal category. Conversely speakers 003, 064 and 040 shifted into the Dove category at the lowest frequency bandwidth 0-3kHz, 06kHzSR from the normal position.

## 10.6 Responses to Questions

**Q1 Recap** Does ASR performance noticeably improve relative to baseline when the frequency bandwidth is extended beyond telephony? If so, what is the optimum frequency bandwidth for ASR performance?

**A1** As predicted and in line with research, performance for both ASR systems improved for wide band speech in comparison to narrow band speech. However, the performance differential was relatively marginal. This supports the hypothesis that, whilst some discriminatory speech information does extend beyond the 4kHz frequency point, the majority of speech information (F1, F2, F3) for successful ASR discrimination occurs within the 0-4kHz frequency bandwidth (sample rate of 8kHz). In line with prediction, the i-vector ASR system had a closer correlation of optimum settings to the bandwidth of speech under matched conditions (0-11kHz, SR 22kHz).

In terms of the bespoke normative data and settings, it was difficult to assess how well the ASRs were optimised for wide band use and whether results could be improved on further. Further research is required to validate that the optimised range for iVocalise could be replicated on other types of ASR systems. Many ASR systems do not have options to optimise them for usage beyond narrow band/telephony.

To restate, the experiments were carried out solely on male speech data. It is likely that extending the frequency bandwidth for female/child speech data is likely to offer additional ASR performance benefit, assuming that the normative data reflected the demographic changes. Further research is recommended.

**Q2 Recap Does an i-vector/UBM, TV, LDA+PLDA ASR system offer significant performance advantages over a GMM-UBM system when the frequency bandwidth is extended?**

**A2** Yes. As anticipated the i-vector/PLDA system outperformed the GMM-UBM system in all conditions across all performance metrics. The i-vector version was also more consistent in output with less variability in EER% under matched conditions. Optimum performance varied between the two systems under matched conditions.

- i. iVocalise, 0-11kHz, (22kHz SR) = 0.027 EER%. 0-4kHz (8kHz SR) = 0.042 EER%
- ii. Vocalise, 0-11kHz (22kHz SR) = 0.27 EER%. 0-4kHz (8kHz SR) = 2.43 EER%

The accuracy (Cllr) of both systems was also examined. The i-vector system consistently outperformed the GMM-UBM system under matched conditions.

- i. iVocalise, 0-11kHz, (22kHz SR) = 0.02 Cllr, 0-4kHz (8kHz SR) = 0.01 Cllr
- ii. Vocalise, 0-14kHz, (28kHz SR) = 0.86 Cllr, 0-4kHz (8kHz SR) = 0.48 Cllr

As predicted, performance improvements were noted for the i-vector/PLDA system under matched conditions when the frequency bandwidth was extended (i-vector 0-11kHz, 22kHz SR = 0.027 EER% vs WB GMM-UBM 0-14kHz, 28kHz SR = 0.27 EER%). Improvements in Cllr/accuracy were also noted supporting the hypothesis that better underlying statistical modeling in the i-vector ASR system enables better exploitation of acoustic data containing more information, better EER% performance and higher system accuracy. Of further note, true positive likelihood scores rose and true negative scores fell demonstrating an improvement of score separation. Good score separation is important both for setting system thresholds and for assisting with interpreting results.

**Q3 Recap Many ASR systems automatically downsample audio files as they are imported, to a frequency bandwidth 0-4kHz (sample rate 8kHz). OWR Vocalise and iVocalise ASR software systems provide the operator with the opportunity to adjust the frequency bandwidth (minimum and maximum settings) for the MFCC feature extraction stage and allow the configuration of normative data. Can performance advantages therefore be found in terms of matching frequency bandwidth for speaker models and test audio? how would ASR performance vary against baseline?**

- **If iterative bandwidth degradation was applied to the test audio but wide band speaker models were used, how would ASR performance vary against baseline?**

**A3** As hypothesised, performance advantages were observed when matching the channel bandwidth of SM and TA. Conversely, the performance noticeably deteriorated for the lowest frequency bandwidth comparisons for unmatched conditions in both systems. For the i-vector system the performance differential between matched and unmatched conditions was more significant at the lowest frequency bandwidth settings.

- Matched: 0-3.5kHz (SR 7kHz) = 0.33 EER% and 0-3kHz (SR 6kHz) = 1.01 EER%.
- Unmatched: 0-3.5kHz (SR 7kHz) = 16.30 EER% and 0-3kHz (SR 6kHz) = 43.8 EER%

This is likely due to high divergence between SM and TA affecting statistical modelling and therefore comparison (i.e. loss of F4 and deterioration in F3 occurring in the TA but not in the SM).

**Q4 Recap      If the frequency bandwidth is significantly reduced below that of standard telephony what implications would that have for ASR performance?**

**A4** As hypothesised, reducing the frequency bandwidth below standard telephony was shown to degrade performance on both systems. A tipping point was observed in both GMM-UBM and i-vector/PLDA systems below 0-4kHz frequency bandwidth (sample rate 8kHz) in unmatched conditions. This is likely due to the degradation in F4 and some elements of (high) F3 as the frequency bandwidth is constrained below the 4kHz frequency point.

For the i-vector system, performance improvements could be observed when extending the frequency bandwidth but only under matched conditions. It was interesting to note the relatively high EER performance for the i-vector system on 0-3kHz frequency bandwidth data, 06kHz SR, obtained under matched conditions (Table 10.14). This suggests that, to some extent performance can be preserved, despite the absence of speech data, if frequency bandwidth conditions are better matched between SM and TA. This is likely due to the richness/density of the statistical model (i.e. i-vector).

**Table 10.14:** Comparison in low bandwidth speech, EER% (i-vector/PLDA ASR)

I-vector/PLDA ASR (matched SM/TA)	I-vector/PLDA ASR (unmatched SM/TA)
0-3.5kHz, SR7kHz = 0.33 EER%	0-3.5kHz, SR7kHz = 16.30 EER%
0-3kHz, SR6kHz = 1.00 EER%	0-3kHz, SR6kHz = 43.80 EER%

### **10.6.1 Summary of Results**

Experiments demonstrated that ASR system performance degraded when frequency bandwidth constraints were applied and that was more noticeable in unmatched conditions (SM and TA). In addition, degradation accelerated as the lowest frequency bandwidth settings were reached.

It was shown that an i-vector ASR system can provide performance and accuracy dividend over a GMM-UBM system and that matched conditions consistently performed better for both ASR types.

Results confirmed that the optimum frequency bandwidth settings broadly reflected that of speech with 0.0269 EER% obtainable at 0-11kHz, SR 22kHz (iVocalise system under matched conditions).

As stated, DyViS data features only male speakers. Male speakers with higher F3 and F4 mean values are likely to be more affected by constraining the frequency bandwidth below telephony than those with lower F3 and F4 mean values. Results suggest that using the same ASR, with frequency bandwidth set to 0-4kHz, to compare female speakers would perform marginally worse since their speech generally contains higher average formant (F1, F2, F3) frequencies. Although also not tested - child speech would be likely to perform worse than females.

It was also demonstrated that when extending the frequency bandwidth, on both ASR systems, performance benefit (EER% and Cllr) could be gained over standard telephony bandwidth. Benefit was greater in the i-vector system for extended frequency bandwidth under matched conditions.

## **10.7 Discussion and Practical Application**

This section discusses the broader implications of the results from the experiments and places them into a practical context (e.g. investigative and forensic casework). It provides recommendations based on findings in reference to the thesis objectives.

At a practical level, results support that frequency bandwidth should be examined and considered at the technical assessment stage, i.e. prior to ASR analysis, preventing the use of ASRs on unsuitable audio. Wideband ASRs are likely to perform better than narrowband ASRs, particularly with regard to female and child speech. ASR systems should be kept up to date in order to benefit from advances in technology/statistical modelling. Results also tentatively support the hypothesis that there currently may be little dividend in extending sample rates for speech recordings beyond 22kHz (solely in the context of ASR/i-vector analysis).



On listening to approximately 20% of the audio files (>4kHz), most of the intelligible speech was gone and auditory discrimination would be next to impossible. Nevertheless, within the high frequency unintelligible whispers, speakers exhibited slightly different qualities and certain speakers sounded dissimilar.

Channels of communication are constantly evolving and casework requirements can arise where it is tempting to apply ASR analysis to speech data irrespective of frequency bandwidth. In casework it is conceded that matching conditions for SM and questioned audio is likely to be impossible. Nevertheless, where the questioned audio exceeds the frequency bandwidth of the speaker model(s), results support the hypothesis that it might be possible to complete a controlled sample rate conversion ensuring the use of a brick-wall LPF to bring greater parity to conditions. Whilst this is likely to be controversial, due to the unmeasurable affect that this would have on case data, results showed it would be more likely to improve ASR performance than on an unmatched comparison.

It was observed that performance did not degrade significantly by incorporating frequencies at the very high end of the frequency spectrum (i.e. >12kHz). The DyViS samples are well recorded and the general lack of noise in the >12kHz frequency range notable. Therefore, extending the frequency bandwidth to a very high frequency would be unlikely to register performance degradation in the experiments. Whilst untested, a further recommendation would be in applying a LPF (at approximately 12kHz) if considering upgrading a narrow band ASR to a wide band system, simply to avoid any non-speech noise contamination at very high frequencies.

Performance differences were found between GMM-UBM and i-vector/PLDA systems. As the frequency bandwidth dropped to 0-3kHz, SR 6kHz the EER elevated to 5.69% on the GMM-UBM system in comparison to 1.0067%. Performance gains could therefore be achieved by upgrading an older GMM-UBM ASR to an i-vector/PLDA system, especially when conducting speaker comparisons on lower frequency bandwidth samples (<0-4kHz). In terms of practical recommendations, the experiments again highlight the importance of ensuring ASR systems are up to date.

Orman and Aslan (2001) suggested a revised filter bank that improved on the Mel scale, stating it equally as important to improve the feature extraction part of the process as to improve the modelling. Using the animated zoo plots it was also observed that the score height generally fell as the band limitation increased, even when the EER% was not particularly affected. This was evidenced by the overall data points moving from the top right quadrant to the lower left. LLR Score trends like this are important to be aware of when examining a mixture of audio files at different sample rates. It is also useful for setting system thresholds to mitigate for relatively low true positive or high false positive scores. It was shown that in unmatched conditions, where either

the speaker model or test/questioned audio is of a higher sample rate, applying good quality sample rate conversion prior to ASR ingest should realise a performance benefit in both EER% and Cllr for either i-vector or GMM-UBM systems. Finally, a tipping point or cliff effect was visible suggesting a potential threshold for ASR application. This was particularly evident on the GMM Vocalise system and occurred as the frequency bandwidth dropped below 0-4kHz, 8kHz SR.

In low net duration speech samples which are also frequency bandwidth constrained, it is suggested that this could mean that it would become much more important as to what was said and how phonetically rich that data is – within that constraint.

Finally, on referring back to the preliminary testing and LTFD trials, it was noted that the difference in EER% from extending the LTFD extraction from F1, F2, and F3 to include F4 also provided a small performance improvement (Table 10.15, and 6.1). This also supports the importance of higher frequencies (to ASR) and clearly, the inclusion of F4, assists with performance through the provision of more speaker-discriminatory information.

**Table 10.15:** Results from preliminary tests, showing influence on EER% re addition of F4

Software	Engine	UBM	Extraction settings	EER %
Vocalise	LTFD	Type A SSBE UBM	F1, F2, F3 32Gaussians	7.483
<u>Vocalise</u>	<u>LTFD</u>	<u>Type A SSBE</u> <u>UBM</u>	<u>F1, F2, F3, F4 32 Gaussians</u> <u>(optimum)</u>	<u>6.022</u>

# Chapter 11 Transcoding

---

## 11.1 Introduction

This chapter examines the influence of transcoded speech files on ASR systems, completing ten experiments (including baseline) under controlled conditions to determine the extent to which different codecs can degrade ASR performance.

Baseline performance tests were created from one hundred DyViS speakers in .wav PCM format (task 1, mock interview data). These were edited to provide 100 speaker models (SM) and 300 test audio (TA) files. Nine different codecs were then applied to the baseline data using a total of 53 different data compression settings.

The baseline and transcoded data was then examined using two different ASR systems a GMM-UBM and an i-vector UBM, TV, LDA+PLDA. Experiments were completed under matched speaker model (SM) and test audio (TA) conditions. From preliminary experiments and research completed by others, it was established that mismatched conditions would provide poorer ASR performance. In addition, permutations of mismatched conditions are almost infinite. Only matched conditions were therefore considered in scope. Results were analysed with regard to the transcoded material and baseline/control data (non transcoded).

The chapter begins with a review of related research. Questions are then specified with associated hypotheses. An outline is presented of the experiments completed and results provided. The research questions are revisited and the chapter concludes with discussion, offering several practical recommendations for approaching transcoded casework data using ASR systems.

## 11.2 Background

To store, transmit and receive speech digitally requires a coder-decoder algorithm, commonly referred to as a codec.

The UK has seen a transition from traditional landline and mobile telecommunications channels to integrated telecomms and computer network systems. With the upsurge in smartphone use, upgraded 4G infrastructure, broadband and wifi methods of communications there has also been an increase in the types of codecs used and numbers of transcoding steps. Higher data transfer speeds ensure audio and video exchanges are fast and generally higher in quality than traditional narrowband methods of communication.

Examples include voice over internet protocol (VoIP) events, audio material from social media sites or speech data from smartphone applications. Transcoding algorithms which were previously more often used in IT systems are now encountered in telecommunications channels (e.g. GSM or Speex/Opus). In addition, codecs can combine in series as speech transitions through telecomms and IT infrastrucutre. As the signal path of the speech is more often unknown, it can become difficult to accurately assess which codecs speech has passed through and therefore to analyse speech accurately. A greater variety of codecs integrated in the signal path can degrade the technical quality of speech in ways which can be difficult to quantify by a forensic examiner.

Some codecs are regarded as lossless whilst others employ constrained and variable bit rates in terms of kilobytes per second (kbps) to preserve data bandwidth. This can cause a codec to adapt compression levels to changing broadband/network speeds which then produces variable data compression and/or frequency bandwidths. Other types of corruption can occur too such as subsecond data/packet loss, buffering and glitches or interference.

The motivation behind this research was to measure and examine the extent of degradation caused by a selection of codecs and to determine the degree to which ASR performance was affected.

### 11.3 Literature Review

Previous research has examined transcoding degradation in relation to ASR performance.

Polacký, Pocta and Jarina (2016a; 2016b) describe codec degradation as one of the most prominent issues relating to telecommunications networks. Their experiments used the TIMIT corpus (Linguistic Data Consortium) to assess 5 different codecs using a GMM-UBM ASR. The codecs tested were G.711.1 at 96kbps, G.729 at 32kbps, AMR-WB, EVS-WB and Speex at 27.8kbps. The term wide band (WB) generally refers to codecs operating at approximately 14kHz sample rate, as opposed to narrow band (NB) for those operating up to 7kHz sample rate. The results were that EER% rates did not fluctuate significantly from a statistical context but were consistently better for matched rather than mismatched conditions. Increasing compression rates degraded performance. Speex and enhanced voice services (EVS), for 4G, codecs performed well although the quality settings for the former vary in terms of data compression rates and were not specified. In all instances mismatched conditions performed poorer than matched conditions.

**Table 11.1:** Polacký, Pocta and Jarina EER% results (2016a: p.81)

	<b>G.711.1</b>	<b>G.729.1</b>	<b>AMR 6.6</b>	<b>AMR 8.85</b>	<b>AMR 12.65</b>	<b>EVS 5.9</b>	<b>EVS 8</b>	<b>EVS 13.2</b>	<b>Speex</b>
Unmatched	4.37	4.11	10.74	8.22	6.85	8.95	7.01	2.26	2.45
Matched	4.11	3.16	3.68	3.25	3.21	3.48	3.26	2.55	2.43

Jarina, Polacký, Počta and Chmulík (2017) expanded on the research using TIMIT and a GMM-UBM ASR system to examine the influence of VoIP on performance. Their results showed that G.711 and EVS, at higher settings, produced consistently better ASR performance than other codecs tested. However, it should be noted that the constructed UBM contained all the speakers from the TIMIT database which could artificially improve performance level.

Silovsky, Cerva and Zdansky (2011) evaluated 11 different lossy codecs in common use at the time. This is reproduced below in Table 11.2 with their results presented in Table 11.3. Codecs were assessed against baseline data using an unspecified GMM-UBM ASR system and a corpus of 273 speakers in spontaneous telephone conversation to generate SM and TA. Their experiments were set for ‘matched’ where SM and TA were both passed through the codec (ideal conditions) and ‘mismatched’ with only the TA passed through the codec (non-ideal conditions).

In Table 11.2 (below) the following abbreviations apply.

- DTX/CNG. Discontinuous Transmission with Comfort Noise Generation. Transmission is switched off and noise, relevant to the background noise during silent sections of speech, is generated to fill otherwise ‘empty’ sections of conversation. This provides a more fluid communications experience to the participants and ensures that the impression is not given that the call has ended.
- PLC. Packet Loss Concealment. If a piece of transmitted information is missing then lost speech frames can, for example, be replaced by repeating a portion of the waveform or interpolating between the successfully transmitted sections.
- VBR. Variable Bit Rate. Dependent on the data bandwidth available the codec bit rate adjusts accordingly.
- MOS. Mean Opinion Score is effectively a score (1 = poor to 5 = good) which represents the perceived quality of the signal after compression and/or transmission. MOS-ic and MOS-ns refers to ideal conditions or network stress accordingly (as defined by ITU-T standards).

**Table 11.2:** Silovsky, Cerva and Zdansky codecs evaluated (2011: p.206)

Codec	Creator	Supported bitrates [kb/s]	Algorithm	DTX/CNG	PLC	VBR	MOS-ic	MOS-ns
G.711 A-law	ITU-T	64.0	log. PCM	yes	yes	no	4.45	4.11
G.726	ITU-T	16.0 / 24.0 / 32.0 / 40.0 <sup>a</sup>	ADPCM	no	no	no	4.3 @ 32 kb/s	3.79 @ 32 kb/s
G.728	ITU-T	16.0	LD-CELP	no	no	no	N/A	N/A
G.729 annex I	ITU-T	6.4 / 8.0 / 11.8	CS-ACELP	yes	no	no	4.04 @ 8 kb/s	3.51 @ 8 kb/s
G.723.1 annex A	ITU-T	6.3	MPC-MLQ	yes	no	no	4.08	3.57
G.723.1 annex A	ITU-T	5.3	ACELP	yes	no	no	3.65 <sup>b</sup>	N/A
GSM-FR	ETSI	13.0	RPE-LTP	no	no	no	3.5 <sup>b</sup>	N/A
GSM-HR	ETSI	5.6	VSELP	yes	no	no	N/A	N/A
AMR	3GPP	4.75 / 5.15 / 5.9 / 6.7 / 7.4 / 7.95 / 10.2 / 12.2	ACELP	yes	yes	yes	4.15 @ 12.2 kb/s	3.79 @ 12.2 kb/s
iLBC	Global IP Solutions	13.33 / 15.2	BI-LPC	no	yes	no	4.14 @ 15.2 kb/s	N/A
Speex	Xiph.Org Foundation	2.15 – 24.6	CELP	yes	yes	yes	N/A	N/A
SILK	Skype	6.0 – 20.0	LP	yes	yes	yes	N/A	N/A
<sup>a</sup> Bitrate 40 kb/s is not intended for speech encoding and transmission		<sup>b</sup> This value is not relative to the reference MOS value 4.45 for the G.711 codec						

**Table 11.3:** Silovsky, Cerva and Zdansky results (2011: p.207) + annotation

Codec	Bitrate [kb/s]	Mismatched condition		Matched condition	
		EER [%]	Rel.drop [%]	EER [%]	Rel.drop [%]
BASELINE	64.0	7.74	0.00	7.74	0
G.726	16.0	11.41	47.42	11.02	42.38
	24.0	8.90	14.99	9.18	18.6
	32.0	7.77	0.39	7.48	-3.36
G.728	16.0	7.91	2.20	7.76	0.26
G.729	6.4	9.04	16.80	9.89	27.78
	8.0	8.19	5.81	8.19	5.81
	11.8	8.22	6.20	7.23	-6.59
G.723.1	5.3	9.46	22.22	10.05	29.85
	6.3	8.47	9.43	9.19	18.73
GSM-FR	13.0	8.19	5.81	9.32	20.41
GSM-HR	5.6	9.43	21.83	10.62	37.21
AMR	4.8	9.48	22.48	10.03	29.59
	5.2	9.34	20.67	9.75	25.97
	5.9	8.60	11.11	9.18	18.6
	6.7	8.36	8.01	8.62	11.37
	7.4	8.47	9.43	8.91	15.12
	8.0	8.74	12.92	8.62	11.37
	10.2	8.21	6.07	7.24	-6.46
	12.2	8.33	7.62	7.49	-3.23
iLBC	13.3	8.04	3.88	8.31	7.36
	15.2	8.05	4.01	7.34	-5.17
Speex	4.0	10.88	40.57	12.26	58.4
	8.0	9.06	17.05	9.59	23.9
	15.0	7.77	0.39	7.77	0.39
SILK	5.0	8.90	14.99	9.32	20.41
	8.0	7.92	2.33	7.77	0.39
	15.0	7.34	-5.17	7.2	-6.98

 Indicates performance improvement (relative to baseline)

Silovsky, Cerva and Zdansky (2011) results demonstrated that ASR performance drops when applying almost all codecs and in particular Speex and/or those with relatively low kbps settings (e.g. G.726 16kbps). The small EER% performance improvements were not explained, but a plausible explanation could relate to matched SM and TA conditions and/or the composition of the normative data with respect to codec. The main purpose of their experiment was to examine telecommunications speech data and the specification of the untreated audio was therefore 8 bit A-law and sampled at 8kHz. This could explain their relatively high baseline of 7.74% EER on a GMM-UBM ASR. It was noted that some speech samples were less than 10 seconds in duration and it is suggested that net duration was likely to have also influenced results as in chapter 7. SILK

provided less degradation overall and in one instance marginally improved against baseline results (-5.17% EER). In conclusion, it was found that most codecs negatively influenced ASR performance with a few exceptions that could relate to closer matching of channel conditions and/or normative data with respect to codec. This paper assisted in informing the methodology for the experiments conducted in this chapter and added further objectives such as using wider band speech, with greater net duration and a more modern i-vector/UBM, TV, LDA+PLDA system.

Janicki and Staroszczyk (2011) used a GMM-UBM ASR system and the TIMIT corpus of 630 speakers to examine the effects of 6 codecs on performance. The codecs assessed, predominately used in telecommunications at the time, were G.711, G.723, GSM06.10, GSM06.60, G.729 and Speex setting 8. It was interesting to note that the range of codecs has diversified over recent years to include a much wider variety of propriety codecs such as 3GPP, Opus and ADPCM. Widening the range of codecs in use is likely to make ASR benchmarking and subsequent analysis more difficult since the technical influence of each codec is unknown. One of the key objectives of the study was to find which codec created speaker models that were the most resilient to mismatch using support vector machine classification (SVM). The utterances in TIMIT are relatively short (3.2s average) suggesting that net duration was likely to have influenced results (chapter 7) but they were relatively higher in quality than those used in the Silovsky study at 16bit, 16kHz sample rate suggesting that the increase in frequency bandwidth could potentially offset the low net duration. Results were expressed as percentage correct rather than EER%.

**Table 11.4:** Janicki and Staroszczyk codec results (2011: p.296), bold=best % correct

training/testing	un-coded	G.711	G.723	G.729	GSM 06.10	GSM 06.60	Speex 8	average	stddev
uncoded	<b>89.67</b>	87.40	49.49	51.49	51.49	51.44	83.63	66.37	17.59
G.711	86.42	<b>88.23</b>	46.98	47.02	50.47	64.65	80.60	66.34	16.07
G.723.1	71.63	68.88	<b>73.81</b>	61.63	60.33	71.30	75.58	69.02	4.64
G.729	65.16	62.37	57.95	<b>77.12</b>	37.72	77.91	62.74	63.00	8.91
GSM 06.10	69.63	70.98	55.26	44.98	<b>83.12</b>	57.72	72.23	64.84	10.45
GSM 06.60	72.00	65.54	63.07	63.21	41.86	<b>84.28</b>	63.07	64.72	7.90
Speex 8	86.60	84.23	62.88	53.07	62.79	67.67	<b>86.65</b>	71.99	11.87

Janicki and Staroszczyk (2011) demonstrated that ideal/matched conditions outperformed the mismatched conditions. Conversely to the Silovsky study, Janicki and Staroszczyk suggested that Speex provided less deterioration in performance against baseline. Again, the difference in findings could be explained by codec configuration, since Speex has 15 different modes with 10 quality settings and the full detail of the exact configuration was not provided.



Another interesting finding in the Janicki and Staroszczyk (2011) study was that the increase in the number of gaussian mixture models (i.e. improving the detail/richness in the statistical model for the degraded speech files over the baseline number of gaussians extracted) appeared to also benefit ASR performance (2011: pp.295-297).

Nandan and Saha (2012) examined bit rate, noise addition and packet loss in the context of VoIP and mobile communication in relation to ASR performance (GMM-UBM). They used the YOHO corpus (1994) which features English language speakers reading aloud two digit numbers. It is recorded at 4kHz bandwidth, microphone, single channel pcm. Nandan and Saha showed that the performance loss caused by GSM-AMR was important, was relatively large 2.35% to 12.2% EER and noise also degraded performance (2.35% EER to 11.22% EER at 20db SNR). Nandan and Saha suggested that packet loss was, overall, somewhat less important to ASR performance although their detail is not provided as to the extent. Nandan and Saha also make an assumption that 'lowering the bit-rate does not compromise with the speaker's 'biometric identify' (2012: p.4). This runs counter to the experiments completed in this thesis (chapter 11). It is also suggested that net duration likely influenced their performance figures (see chapter 7).

Becker, Broß and Meier (2011) examined MP3 compression on a bespoke ASR system (GMM-UBM) using 102 male Romanian speakers recorded at 8kHz, 16bit. They found a significant deterioration in performance at very low conversion bit rates to MP3 (8kbps) although EER% in standard terms is not expressed. However, they also found that compression at other rates (16kbps and 32kbps) actually caused some improvement in discrimination performance for certain recordings, in terms of TP LLR score size and separation from FP. The study also pointed out that transcoding history is often an unknown variable.

In the Polacký, Pocta and Jarina (2016) study marginal degradation of ASR performance was described as nonsignificant (EER% ranged from 2.43% to 10.74%). In their study, the poorest performing ivector system was 2% EER against a baseline of .0051%. It could be argued that a system that is 98% accurate as opposed to 99.99% accurate is unimportant. However, other factors, besides EER%, must be taken into consideration than simply the transcoding of high quality speech data from a well recorded corpus such as accuracy (Cllr) and score separation (LR/LLR).

Research by Petracca, Servetti and DeMartin (2006) examined several codecs (and net duration). They used a GMM-UBM system to examine GSM AMR, G.729, G.723 and two other proprietary VoIP codecs at 10 different settings. The dataset used consisted of 14 speakers which, it could be argued, is a relatively small set of data. The TA length was varied in relation to the SM (10s, 20s and 30s). The group concluded that the 'recognition performance does not always decrease with the coder bit rate' (2006, p.1396). This is true, but to be expected, since codecs work in different

ways in terms of compression on different aspects and ranges of the frequency spectrum. Interestingly, their research also demonstrated that increasing net duration could offset ASR performance loss from codec compression (to some extent).

Stauffer and Lawson (2009) studied the Speex codec at different settings on an unspecified GMM-UBM ASR system using a bespoke corpus of 240 conversations at 120 seconds long, 60 second net at 8kHz sample rate, 16bit. Their study found that Speex, as long as it was applied at the highest quality settings, actually produced only a 1% drop in relative (A)SR performance although the low quality settings degraded performance by 22% (2009: p.2366).

## **11.4 Questions and Hypotheses**

After consideration of research literature the following research questions were defined and hypotheses formulated.

**Q1 How resilient are more modern i-vector/PLDA ASR systems to codec degradation in comparison with GMM-UBM systems?**

**H1** I-vector/PLDA systems are likely to be more resilient to the loss of speech data through compression than GMM-UBM systems. This is due to the improvements in the accuracy of statistical modelling as well as other modifications to the feature extraction and speech detection phases.

**Q2 To what extent does ASR performance degrade when transcoding processes are applied to baseline data?**

**H2** Since transcoding often removes data through compression and/or band limitation ASR performance will be degraded in all cases to varying degrees. This is likely to be proportionate to the extent of the data loss and any band limiting with regard to speech frequencies (approximately >50Hz to 16kHz). Effectively, greater deterioration to the speech formants will result in larger performance loss.

**Q3 How will compression codecs influence ASR performance?**

**H3** Performance decreases are likely to be proportionate to the extent of the data compression inherent in the transcoding settings. More compression will produce greater degradation in EER%. Codecs which also limit frequency bandwidth are also likely to degrade performance, as noted in chapter 10. Codecs which add noise and therefore limit the net duration, which passes the speech detection or VAD are also likely to degrade ASR performance.

**Q4 Can any operating thresholds be extrapolated relating to data compression rates which may assist with informing ASR use?**

**H4** Measuring performance thresholds will be extremely difficult due to the very large variety of codecs in existence, the multiple settings that they use and the variation inherent in speech and recoding environments. However, it may be possible to form some broad conclusions if there are codecs/settings with important deleterious effects on ASR performance under both GMM-UBM and i-vector systems.

## 11.5 Methodology and Materials

Please see chapter 5, with the following adaptations.

Baseline data comprises 100 DyViS speakers from the task 1 interview channel in .wav PCM format. This was edited to generate 100 speaker models and 300 test audio files. These comprised of 1x minute SM and the remaining file divided to create 2x 1 minute TA files with residual data comprising the third test audio file.

### 11.5.1 List of Codecs for Comparison

Nine codecs with a total of 53 different settings were chosen for comparison. These were chosen to pertain to telecommunications and computer network application (Table 11.5).

**Table 11.5:** Codec types used in experiments with settings

Codec Type	Settings	Comments
WAV	16bit, 22kHz	352kbps (Control/Baseline).
Speex	Quality 0 to 10	10 = highest quality
MP3 CBR	8, 16, 32, 64 & 128 kbps	Constant Bit Rate
MP3 VBR	8kbps to 320kbps VBR	Variable Bit Rate. Quality 4 & 9(high)
ADPCM	6, 8, 16, 22kHz	Dialogic
G. 711	6, 8, 16, 22kHz	uLaw
G. 711	6, 8, 16, 22kHz	aLaw
AMR	4.75, 7.4, 12.2kbps	3GPP
Ogg	Quality 0, 1, 2, 3	3 = highest quality
Opus	6, 8, 10, 12kbps	Constant Bit Rate
M4a AAC	10%, 50%, 100%	Variable Bit Rate (% of 120kbps)
GSM	Standard (one setting)	2bit 6.10 Audio Stream (8kHz 16bit)

NCH software, Switch (2017/18 version) ([nchsoftware.com](http://nchsoftware.com)), was used to batch transcode the baseline data through 9 different codecs. For file acceptance into Vocalise and iVocalise it was necessary to transcode back to 16bit PCM .wav (baseline sample rate of 16kHz was applied to avoid conflation of variables). The process was validated using the Free Lossless Audio Codec

[xiph.org/FLAC](http://xiph.org/FLAC)) to ensure that the reconversion back to PCM .wav process did not influence results.

Settings were partially determined by the options available within the conversion software and commonality. Settings were deliberately weighted to favour the lower end of the codec's operating thresholds as very high quality and lossless codecs (e.g. FLAC as used to validate the process) had no effect on performance. Experiments were completed in order, with the lowest codec setting first and incrementally increased until performance was close, or matched, baseline EER% and Cllr (accuracy). Both the SM and TA files were transcoded (matched conditions). As previously stated, mismatched conditions provide poorer performance and were considered out of scope.

A brief summary description of the codec types selected follows. The descriptions refer to the context of these experiments and first generation transcoding (i.e. not passed through any other codec).

**Waveform Audio File or .wav** is a long established (IBM/Microsoft, 1991) lossless audio file format which does not apply data compression. To that extent, wav files are ideal for generating baseline data.

**Speex** [Speex.org](http://Speex.org) was a popular free (open source) data compression codec, last released in December 2016. It has recently been somewhat superseded by Opus (below) although Speex is still prevalent in IT systems and networks.

**MP3 or MPEG3** (Moving Picture Experts Group) [Mpeg.chiariglione.org](http://Mpeg.chiariglione.org) is a lossy compression algorithm that works on psychoacoustic or perceptual principles. In essence, if a human cannot perceive a frequency because it is out of range or obscured by another (louder/more dominant) sound, it is reduced or removed using data compression.

**Dialogic** [Dialogic.com/ADPCM](http://Dialogic.com/ADPCM). Adaptive Differential Pulse Code Modulation (ADPCM) is a data compression algorithm that essentially records the difference between samples and adapts according to the scale of the difference. The data compression is applied on recording and the data is decompressed on playback, offering less loss over other algorithms.

**G.711** [itu.int/rec/T-REC-G.711/en](http://itu.int/rec/T-REC-G.711/en) is a telecommunications codec which was developed in 1972. It is a lossy algorithm which effectively compresses and then expands the dynamic range (companding).

**AMR** [3gpp.org](http://3gpp.org) or Adaptive Multi-Rate codec has undergone several revisions since its introduction in 1999. It was designed for speech transmission and reception and is regarded as

lossy due to the application of data compression. In addition, the codec also constrains frequency bandwidth to 8kHz (13bit) which is also then filtered to 200Hz-3.4kHz.

**Ogg** is an open source file protocol [Xiph.org/ogg](http://Xiph.org/ogg) designed to be very configurable. The extent of data compression depends on the incoming file and, in the context of these experiments, a quality setting (NCH Switch software). Average bit rate is then determined by the incoming file and a quality setting 0 (lowest) to 3 (highest).

**Opus** [Opus-codec.org/](http://Opus-codec.org/) (2012) was also designed by Xiph and combined codecs from Skype (SILK) and constrained energy lapped transform (CELT) to improve the quality of speech whilst addressing some of the latency issues inherent in VoIP communications. Whilst a lossy codec, it is generally accepted that Opus maintains clearer and more intelligible speech at lower bit rates than earlier generation codecs. As a result, Opus is rapidly becoming the industry standard.

**M4a** [Mpeg.chiariglione.org/standards/mpeg-4/audio](http://Mpeg.chiariglione.org/standards/mpeg-4/audio) is essentially an MPEG container which holds only audio, rather than audio and video (as per MP4). The audio algorithm can encode in advanced audio codec (AAC) or the Apple lossless audio codec (ALAC). AAC at three different quality settings was chosen, as these are commonly encountered. They are referred to as ‘10%’ or 12kbps, ‘50%’ or 60kbps and ‘100%’ or 120kbps) and the data compression in this codec works in a perceptual way, similar to MP3.

**GSM** [Etsi.org](http://Etsi.org). GSM or Global System for Mobile communications (version 06.10) is a standard digital, mobile telecommunications codec that uses linear predictive coding or LPC (13bit, 8kHz sample rate).

## 11.5.2 Automatic Speaker Recognition Systems

The baseline and transcoded data was examined using two separate ASR systems (Appendix G):

- i. OWR Vocalise, GMM-UBM system: version 1.5.0.1190, MFCC engine, 32 Gaussians, 16 features.
- ii. OWR iVocalise, i-vector/UBM, TV, LDA+PLDA system: version 2.1.0.1366, PLDA set ‘2016A-1024-D-CMS-Large-VAD-NoDyViS-20Apr16’. The TV was set to 400 dimensions, the PLDA was set to 200 dimensions and 10 train cycles.

Note that, as with all other experiments, neither normative datasets used in the experiments contained any DyViS corpus material, to avoid artificially elevating ASR performance results.

Results were compared with respect to the transcoded material and baseline data (non transcoded). Bio-Metrics version 1.8.0.704 was used for graphing and plotting results from the .csv output files

(EER%, Cllr). Explanations for EER, H0, H1, Cllr, FRR, FAR are provided in chapters 3, 3(3.5) and 9(9.5).

## **11.6 Results**

Tables 11.6 and 11.7 below summarise the results from the i-vector/PLDA and GMM-UBM ASR transcoding experiments.

**Table 11.6:** Transcoding results. OWR iVocalise, i-vector/PLDA ASR

Matched conditions ASR: i-vector/PLDA All files passed VAD													
Test	Codec	Settings	EER	EER ▼▲	% Change relative to baseline EER	Clr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1000	FAR, FRR 10,000
1	Baseline	Control 352kbps 22kHz (SR) .wav	0.0051	↔	0.00	0.113	69.980	-49.929	11.942	26.062	0.00	0.00	1.33
2	Speex	Quality 0 at 16kHz (Low)	2.0084	▼	-39,280	20.984	82.348	28.601	8.758	17.451	3.33	17.84	36.35
3	Speex	Quality 1 at 16kHz	1.2609	▼	-24,623	5.967	70.537	-0.284	10.762	21.028	1.33	9.67	22.36
4	Speex	Quality 2 at 16kHz	0.6397	▼	-12,443	1.989	66.939	-16.706	11.888	22.709	0.33	2.33	8.68
5	Speex	Quality 3 at 16kHz	0.6532	▼	-12,707	1.095	66.974	-23.820	12.081	23.146	0.33	2.33	5.00
6	Speex	Quality 4 at 16kHz	0.3081	▼	-5,941	0.322	66.183	-36.587	12.560	24.298	0.00	0.67	2.33
7	Speex	Quality 5 at 16kHz	0.0320	▼	-527	0.148	66.228	-45.304	12.317	25.333	0.00	0.00	0.67
8	Speex	Quality 6 at 16kHz	0.0303	▼	-494	0.146	66.241	-45.502	12.344	25.344	0.00	0.00	0.67
9	Speex	Quality 7 at 16kHz	0.0185	▼	-263	0.106	66.848	-48.871	12.281	25.698	0.00	0.00	1.00
10	Speex	Quality 8 at 16kHz	0.0118	▼	-131	0.106	66.848	-48.876	12.291	25.698	0.00	0.00	1.00
11	Speex	Quality 9 at 16kHz	0.0236	▼	-362	0.105	67.628	-49.702	12.231	25.849	0.00	0.00	1.33
12	Speex	Quality 10 at 16kHz (High)	0.0286	▼	-460	0.105	67.630	-49.703	12.251	25.841	0.00	0.00	1.34
13	MP3	CBR 8kbps	2.0892	▼	-40,864	13.158	71.626	15.782	10.527	18.733	3.67	19.48	39.33
14	MP3	CBR 16kbps	0.3367	▼	-6,501	0.995	67.717	-28.423	11.828	25.256	0.00	1.00	9.01
15	MP3	CBR 32kbps	0.0084	▼	-64	0.112	67.672	-50.363	11.948	26.125	0.00	0.00	2.00
16	MP3	CBR 64kbps	0.0051	↔	0	0.110	69.285	-49.911	12.092	25.994	0.00	0.00	1.00

Test	Codec	Settings	EER	EER ▼▲	% Change relative to baseline EER	Cllr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
17	MP3	CBR 128kbps	0.0051	↔	0	0.114	69.885	-49.949	11.985	26.079	0.00	0.00	0.67
18	MP3	VBR 8-16kbps Quality 4	0.0051	↔	0	0.108	69.288	-49.816	11.966	25.949	0.00	0.00	0.33
19	MP3	VBR 16-32kbps Quality 4	0.0051	↔	0	0.108	69.288	-49.816	11.966	25.949	0.00	0.00	0.33
20	MP3	VBR 8-16kbps Quality 9 (lowest)	0.0219	▼	-329	0.249	67.526	-39.509	12.190	24.173	0.00	0.00	4.02
21	MP3	VBR 16-32kbps Quality 9 (lowest)	0.0185	▼	-263	0.247	67.453	-39.509	12.129	24.133	0.00	0.00	1.33
22	MP3	VBR 32-64kbps Quality 9 (lowest)	0.0404	▼	-692	0.246	67.372	-39.780	12.191	24.172	0.00	0.00	6.67
23	MP3	VBR 64-128kbps Quality 9 (lowest)	0.032	▼	-527	0.244	67.374	-39.815	12.205	24.168	0.00	0.00	5.37
24	MP3	VBR 128-256kbps Quality 9 (lowest)	0.0522	▼	-923	0.275	68.205	-38.752	11.949	23.930	0.00	0.33	4
25	MP3	VBR 160-320kbps Quality 9 (lowest)	0.0522	▼	-923	0.275	68.205	-38.752	11.949	23.930	0.00	0.33	4
26	ADPCM	Dialogic 6kHz	0.0825	▼	-1,517	20.776	90.343	28.134	7.471	17.578	0.00	0.00	6.68
27	ADPCM	Dialogic 8kHz	0.0269	▼	-427	0.383	68.731	-35.748	11.030	24.772	0.00	0.00	0.67
28	ADPCM	Dialogic 16kHz	0.0034	▲	33	0.100	67.830	-50.532	12.169	26.310	0.00	0.00	0.00
29	ADPCM	Dialogic 22kHz	0.0034	▲	33	0.095	68.293	-51.401	12.332	26.444	0.00	0.00	0.00
30	G.711	uLaw 6kHz	0.0488	▼	-857	10.525	84.909	11.156	8.636	18.701	0.00	0.00	1.33
31	G.711	uLaw 8kHz	0.0067	▼	-31	0.110	68.507	-50.100	11.936	26.262	0.00	0.00	0.33
32	G.711	uLaw 16kHz	0.0051	↔	0	0.113	69.980	-49.929	11.942	26.062	0.00	0.00	1.33
33	G.711	uLaw 22kHz	0.0051	↔	0	0.104	69.521	-50.680	11.914	26.154	0.00	0.00	0.33
34	G.711	aLaw 6kHz	0.0387	▼	-658	10.246	84.920	10.569	8.531	18.803	0.00	0.00	0.67
35	G.711	aLaw 8kHz	0.0034	▲	33	0.100	68.052	-50.916	11.954	26.282	0.00	0.00	0.00
36	G.711	aLaw 16kHz	0.0051	↔	0.00	0.103	69.434	-50.623	11.860	26.162	0.00	0.00	0.33
37	G.711	aLaw 22kHz	0.0051	↔	0.00	0.104	69.450	-50.626	11.880	26.137	0.00	0.00	0.68



Test	Codec	Settings	EER	EER ▼▲	% Change relative to baseline EER	Cllr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
38	AMR	3GPP AMR 4.75kbps	0.9983	▼	-19,474	0.670	64.771	-30.814	12.160	24.651	1.00	2.00	6.68
39	AMR	3GPP AMR 7.4kbps	0.6667	▼	-12,973	0.280	64.575	-39.856	12.181	25.580	0.33	1.00	3.33
40	AMR	3GPP AMR 12.2kbps	0.3316	▼	-6,402	0.168	65.770	-45.470	12.147	26.224	0.00	0.67	1.67
41	OGG	OGG Quality 0 (Lowest)	0.3047	▼	-5,874	0.262	70.594	-44.059	11.897	26.515	0.00	0.67	2.84
42	OGG	OGG Quality 1	0.0051	↔	0	0.113	69.980	-49.929	11.942	26.062	0.00	0.00	1.33
43	OGG	OGG Quality 2	0.0067	▼	-31	0.126	69.310	-48.989	12.134	25.891	0.00	0.00	1.35
44	OGG	OGG Quality 3	0.0067	▼	-31	0.116	69.267	-49.263	12.170	25.873	0.00	0.00	1.34
45	OPUS	CBR (Hard) 6kbps	0.3300	▼	-6,370	1.092	69.603	-23.080	11.278	22.870	0.33	0.33	2.33
46	OPUS	CBR (Hard) 8kbps	0.0387	▼	-658	0.179	65.672	-42.337	12.514	24.862	0.00	0.00	1.67
47	OPUS	CBR (Hard) 10kbps	0.0168	▼	-229	0.153	67.154	-44.759	11.809	25.218	0.00	0.00	0.67
48	OPUS	CBR (Hard) 12kbps	0.0421	▼	-725	0.279	69.619	-37.120	12.439	23.696	0.00	0.00	1.00
49	OPUS	CBR (Hard) 14kbps	0.0236	▼	-363	0.213	69.579	-40.700	12.339	24.245	0.00	0.00	0.33
50	OPUS	CBR (Hard) 16kbps	0.0051	↔	0	0.136	67.632	-46.183	12.202	25.441	0.00	0.00	1.00
51	M4a	AAC VBR Quality 10% [12kbps VBR]	0.7323	▼	-14,259	6.405	73.735	-0.037	9.532	22.141	0.57	3.67	9.70
52	M4a	AAC VBR Quality 50% [60kbps VBR]	0.0236	▼	-362	0.139	69.245	-48.510	11.90094	26.137	0.00	0.00	2.01
53	M4a	AAC VBR Quality 100% [120kbps VBR]	0.0051	↔	0	0.102	69.012	-50.382	11.86721	25.815	0.00	0.00	1.33
54	GSM	2bit 06.10 Audio Stream	0.0051	↔	0	0.371	69.669	-33.982	11.20481	23.201	0.00	0.00	0.33

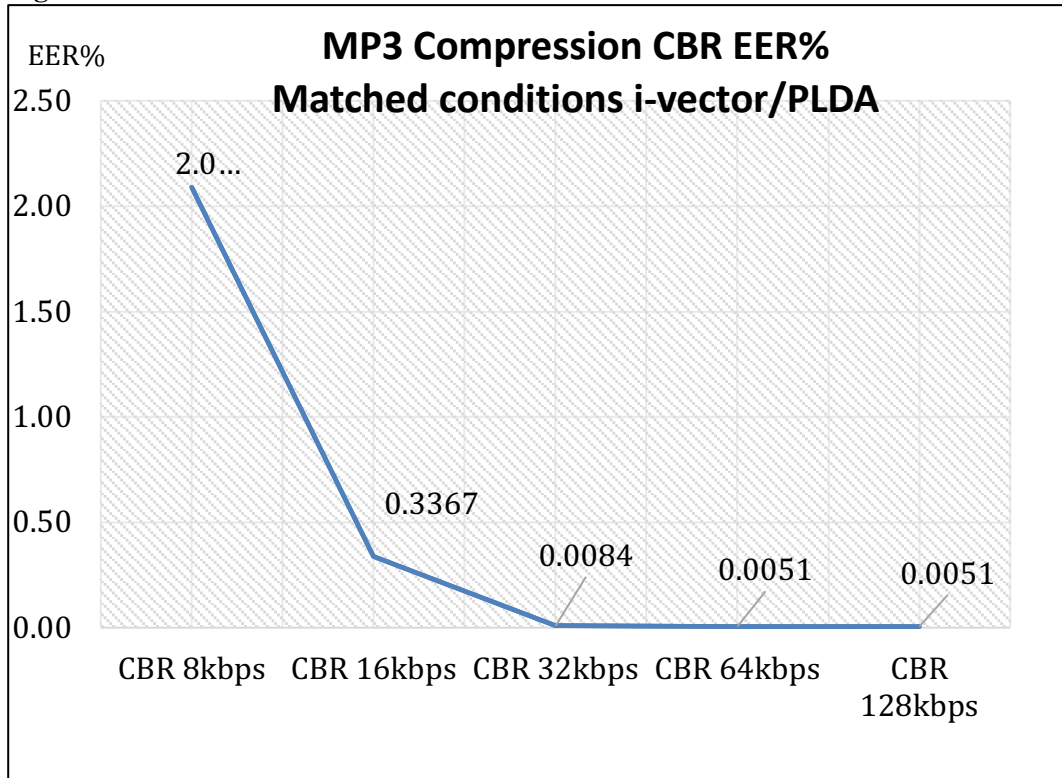
**Table 11.7:** Transcoding results. OWR Vocalise ASR, GMM-UBM

Matched conditions GMM-UBM														
Test	Codec	Settings	HO   H1 Elements (passing VAD)	EER	EER ▼▲	% Change relative to baseline EER	Cllr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR	FAR, FRR	FAR, FRR
												100	1000	10,000
1	Baseline	Control 352kbps 22kHz (SR) .wav	300   29700	2.3889	N/A	0	1.013	3.017	1.008	0.553	0.555	7.00	21.88	53.33
2	Speex	Quality 0 at 16kHz (Low)	300   29700	16.7205	▼	-599	1.861	3.592	2.450	0.654	0.599	64.22	83.00	87.34
3	Speex	Quality 1 at 16kHz	300   29700	15.0522	▼	-530	0.901	1.585	0.470	0.747	0.746	48.53	82.19	92.34
4	Speex	Quality 2 at 16kHz	300   29700	8.2542	▼	-245	0.817	1.764	0.413	0.472	0.580	33.00	65.92	81.00
5	Speex	Quality 3 at 16kHz	300   29700	6.9949	▼	-192	0.806	1.990	0.432	0.504	0.613	30.62	67.33	88.67
6	Speex	Quality 4 at 16kHz	300   29700	5.0589	▼	-111	0.730	2.007	0.251	0.499	0.627	19.04	48.15	71.00
7	Speex	Quality 5 at 16kHz	300   29700	3.7104	▼	-55	0.664	1.996	0.077	0.504	0.635	11.67	32.38	57.35
8	Speex	Quality 6 at 16kHz	300   29700	3.0825	▼	-29	0.643	1.925	0.003	0.484	0.633	9.00	30.53	46.69
9	Speex	Quality 7 at 16kHz	300   29700	3.3569	▼	-40	0.641	2.036	0.021	0.511	0.644	7.39	28.26	43.33
10	Speex	Quality 8 at 16kHz	300   29700	3.6599	▼	-53	0.656	2.092	0.071	0.524	0.652	9.19	31.67	51.00
11	Speex	Quality 9 at 16kHz	300   29700	2.6650	▼	-11	0.657	2.176	0.091	0.528	0.656	7.42	28.22	45.52
12	Speex	Quality 10 at 16kHz (High)	300   29700	2.9209	▼	-22	0.667	2.198	0.120	0.533	0.662	9.67	27.82	43.68
13	MP3	CBR 8kbps	300   29700	28.4933	▼	-1,093	4.830	7.872	6.693	1.115	0.913	81.00	94.67	97.33
14	MP3	CBR 16kbps	300   29700	13.2845	▼	-456	1.342	3.274	1.585	0.791	0.746	57.44	81.07	89.33
15	MP3	CBR 32kbps	300   29700	3.2357	▼	-35	0.633	1.769	-0.054	0.450	0.583	7.63	34.82	64.35
16	MP3	CBR 64kbps	256   25224	3.1404	▼	-31	0.696	2.268	0.217	0.485	0.639	6.71	22.89	41.78

Test	Codec	Settings	HO   H1 Elements (passing VAD)	EER	EER ▼▲	% Change relative to baseline EER	Cllr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR	FAR, FRR	FAR, FRR
												100	1000	10,000
17	MP3	CBR 128kbps	58   5585	4.8602	▼	-103	0.722	2.466	0.305	0.585	0.644	12.07	18.97	36.97
18	MP3	VBR 8-16kbps Quality 4	300   29700	2.2896	▲	4	0.927	2.739	0.823	0.494	0.520	4.77	16.43	40.01
19	MP3	VBR 16-32kbps Quality 4	292   28908	3.0839	▼	-29	0.692	2.311	0.213	0.506	0.645	5.99	23.04	42.88
20	MP3	VBR 8-16kbps Quality 9 (lowest)	230   22702	16.5729	▼	-594	1.328	2.841	1.319	1.311	1.217	66.03	94.48	98.14
21	MP3	VBR 16-32kbps Quality 9 (lowest)	230   22618	14.2988	▼	-499	1.317	2.885	1.371	1.145	1.075	57.67	93.91	99.02
22	MP3	VBR 32-64kbps Quality 9 (lowest)	230   22618	15.6163	▼	-554	1.330	2.826	1.321	1.315	1.195	63.16	93.91	99.57
23	MP3	VBR 64-128kbps Quality 9 (lowest)	230   22618	16.2228	▼	-579	1.309	2.854	1.331	1.183	1.144	60.87	93.94	98.7
24	MP3	VBR 128-256kbps Quality 9 (lowest)	240   23432	17.7948	▼	-645	1.430	3.095	1.615	1.160	0.991	63.31	90.00	97.5
25	MP3	VBR 160-320kbps Quality 9 (lowest)	240   23432	15.7948	▼	-561	1.430	3.095	1.615	1.160	0.991	63.31	90.00	97.5
26	ADPCM	Dialogic 6kHz	300   29700	4.5993	▼	-92	1.210	2.787	1.368	0.418	0.459	13.48	34.63	51.68
27	ADPCM	Dialogic 8kHz	300   29700	1.3266	▲	44	0.587	1.704	-0.197	0.367	0.573	3.00	8.72	20.03
28	ADPCM	Dialogic 16kHz	300   29700	2.0219	▲	15	0.541	1.857	-0.339	0.491	0.688	4.00	15.14	31.34
29	ADPCM	Dialogic 22kHz	300   29700	2.6418	▼	-11	0.627	2.067	-0.013	0.488	0.664	6.00	21.33	33.39
30	G.711	uLaw 6kHz	300   29700	7.9697	▼	-234	1.553	3.662	1.973	0.670	0.586	35.33	72.43	82.67
31	G.711	uLaw 8kHz	300   29700	2.2795	▲	4.6	0.649	2.282	0.101	0.529	0.529	6.00	15.33	30.33
32	G.711	uLaw 16kHz	300   29700	2.9529	▼	-24	0.745	2.532	0.369	0.541	0.666	7.17	23.10	45.34
33	G.711	uLaw 22kHz	300   29700	3.2828	▼	-37	0.724	2.402	0.275	0.605	0.715	6.67	22.77	46.04
34	G.711	aLaw 6kHz	300   29700	8.3064	▼	-248	1.564	3.681	1.991	0.681	0.592	36.57	72.88	85.69
35	G.711	aLaw 8kHz	300   29700	2.5640	▼	-7	0.648	2.294	0.098	0.530	0.626	5.33	14.55	36.69
36	G.711	aLaw 16kHz	300   29700	2.3182	▲	3	0.613	2.275	-0.018	0.543	0.689	4.69	19.67	47.34

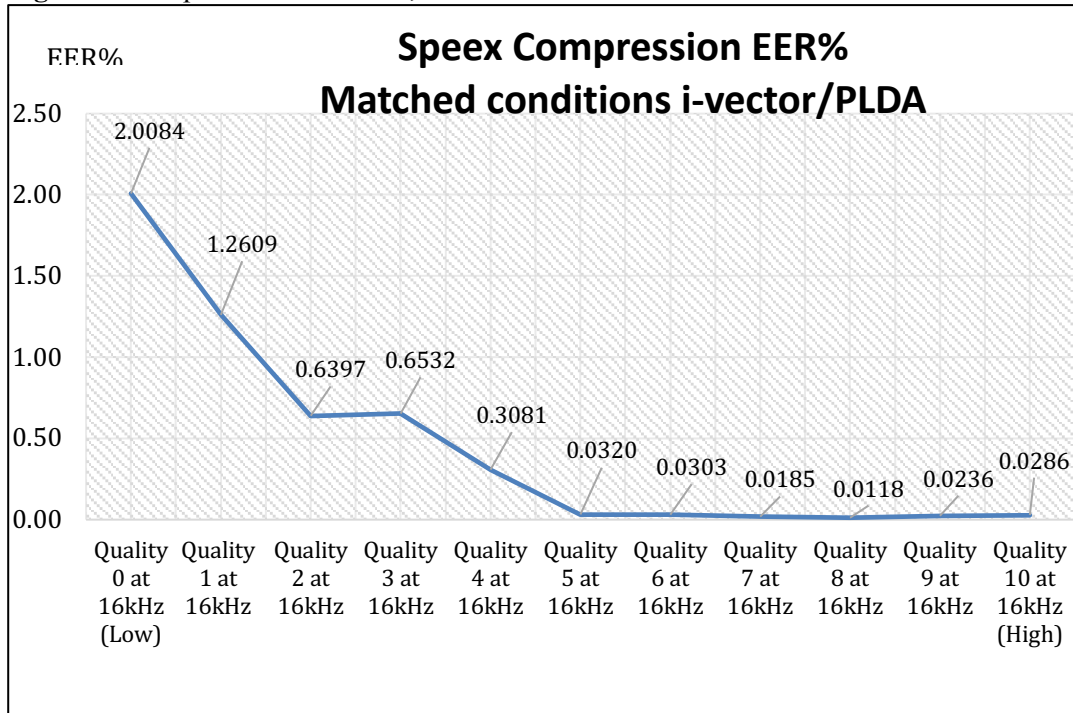
Test	Codec	Settings	HO   H1 Elements (passing VAD)	EER	EER ▼▲	% Change relative to baseline EER	Clr	Mean H0	Mean H1	H0 SD	H1 SD	FAR, FRR 100	FAR, FRR 1,000	FAR, FRR 10,000
37	G.711	aLaw 16kHz	300   29700	2.6549	▼	-11.13	0.715	2.412	0.282	0.517	0.658	5.33	20.33	50.01
38	AMR	3GPP AMR 4.75kbps	300   29700	6.9714	▼	-192	1.248	2.786	1.432	0.467	0.479	25.67	51.62	69.36
39	AMR	3GPP AMR 7.4kbps	300   29700	4.7441	▼	-99	1.141	2.886	1.242	0.487	0.544	16.73	36.33	53.00
40	AMR	3GPP AMR 12.2kbps	300   29700	3.4040	▼	-42	1.037	2.885	1.044	2.885	0.567	10.33	26.67	38.69
41	OGG	OGG Quality 0 (Lowest)	300   29700	4.1044	▼	-72	0.697	2.215	0.198	0.590	0.653	14.90	56.22	78.67
42	OGG	OGG Quality 1	300   29700	2.9529	▼	-24	0.745	2.532	0.666	0.541	0.666	7.17	23.10	45.34
43	OGG	OGG Quality 2	300   29700	3.0034	▼	-26	0.683	2.186	0.170	0.504	0.632	8.33	33.87	63.67
44	OGG	OGG Quality 3	300   29700	2.9680	▼	-24	0.686	2.231	0.184	0.493	0.641	7.14	31.67	49.70
45	OPUS	CBR (Hard) 6kbps - Error	0   0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
46	OPUS	CBR (Hard) 8kbps	300   29700	7.1734	▼	-200	0.773	1.834	0.331	0.556	0.535	34.67	68.67	83.00
47	OPUS	CBR (Hard) 10kbps	300   29700	5.3283	▼	-123	0.726	1.992	0.244	0.578	0.577	24.88	63.87	80.33
48	OPUS	CBR (Hard) 12kbps	300   29700	5.5791	▼	-134	0.750	2.009	0.307	0.586	0.570	27.42	64.72	84.68
49	OPUS	CBR (Hard) 14kbps	300   29700	4.9495	▼	-107	0.723	2.099	0.254	0.607	0.588	22.79	60.00	80.68
50	OPUS	CBR (Hard) 16kbps	300   29700	4.3519	▼	-82	0.723	2.205	0.272	0.595	0.607	19.07	57.22	74.67
51	M4a	AAC VBR Quality 10% [12kbps VBR] Error	0   0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
52	M4a	AAC VBR Quality 50% [60kbps VBR]	300   29700	7.7677	▼	-225	1.044	2.713	1.025	0.577	0.662	42.89	72.36	84.70
53	M4a	AAC VBR Quality 100% [120kbps VBR]	300   29700	3.3451	▼	-40	0.738	2.263	0.323	0.494	0.628	8.94	30.10	62.34
54	GSM	2bit 06.10 Audio Stream	300   29700	2.6448	▼	-11	0.727	1.918	0.263	0.360	0.502	6.33	17.05	30.01

**Figure 11.8:** MP3 EER% results i-vector/PLDA

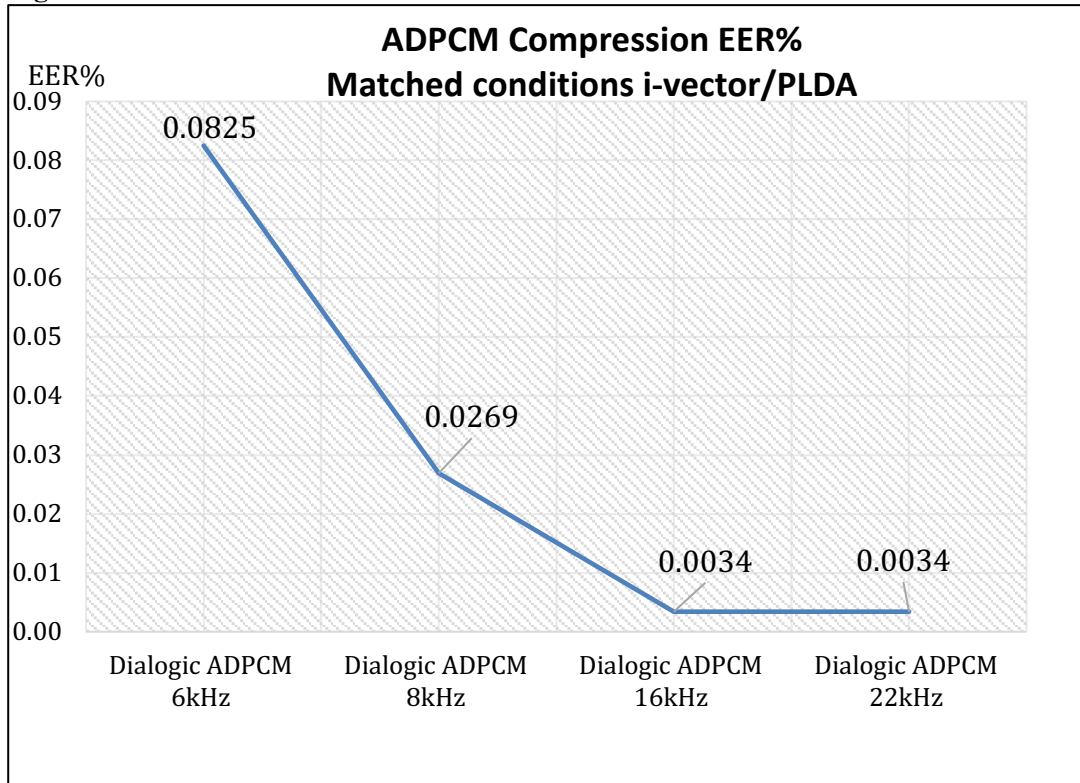


Note the tipping point between 16kbps and 8kbps on Figure 11.8, with diminishing EER% performance gains evident between 32kbps up through to 64kbps and no further benefit up to 128kbps.

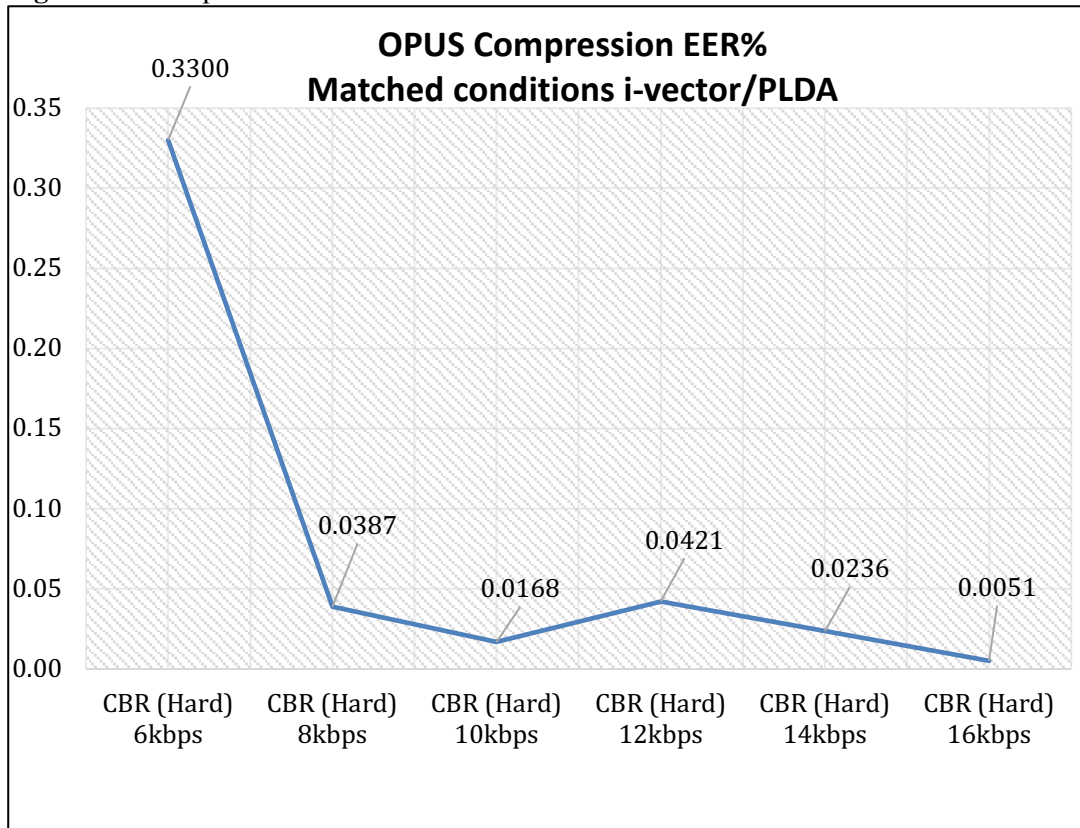
**Figure 11.9:** Speex EER% results, i-vector/PLDA



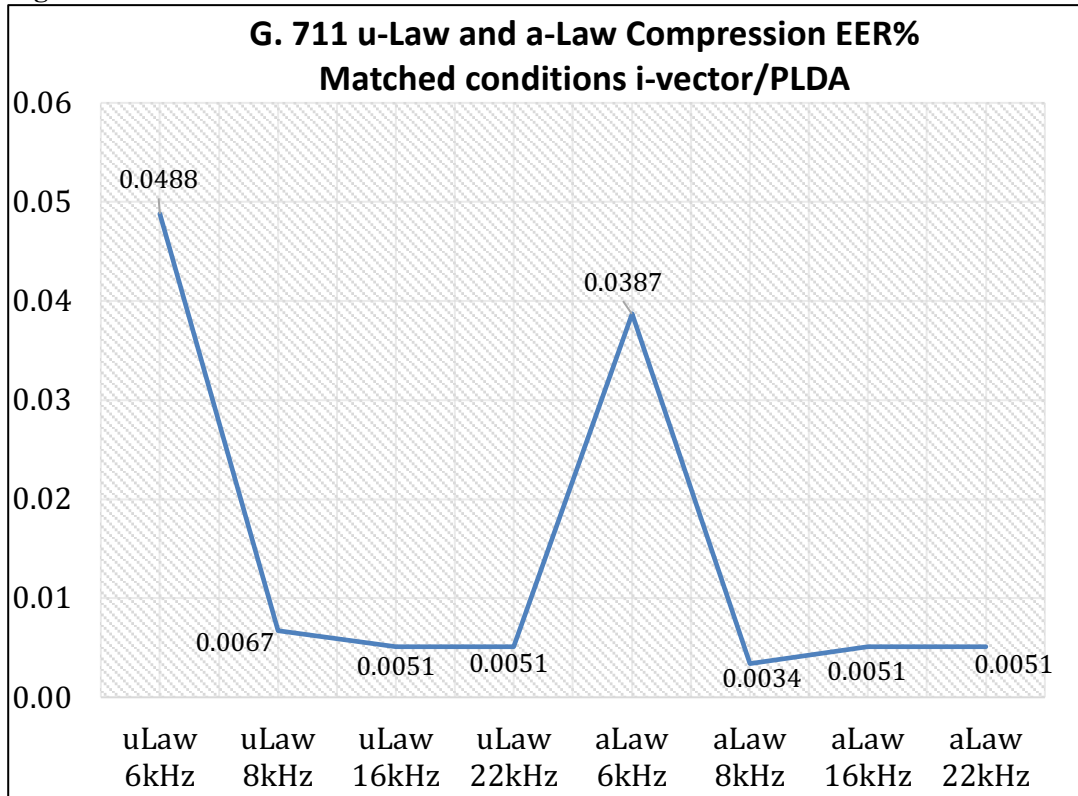
**Figure 11.10:** ADPCM EER% results i-vector/PLDA



**Figure 11.11:** Opus EER% results i-vector/PLDA

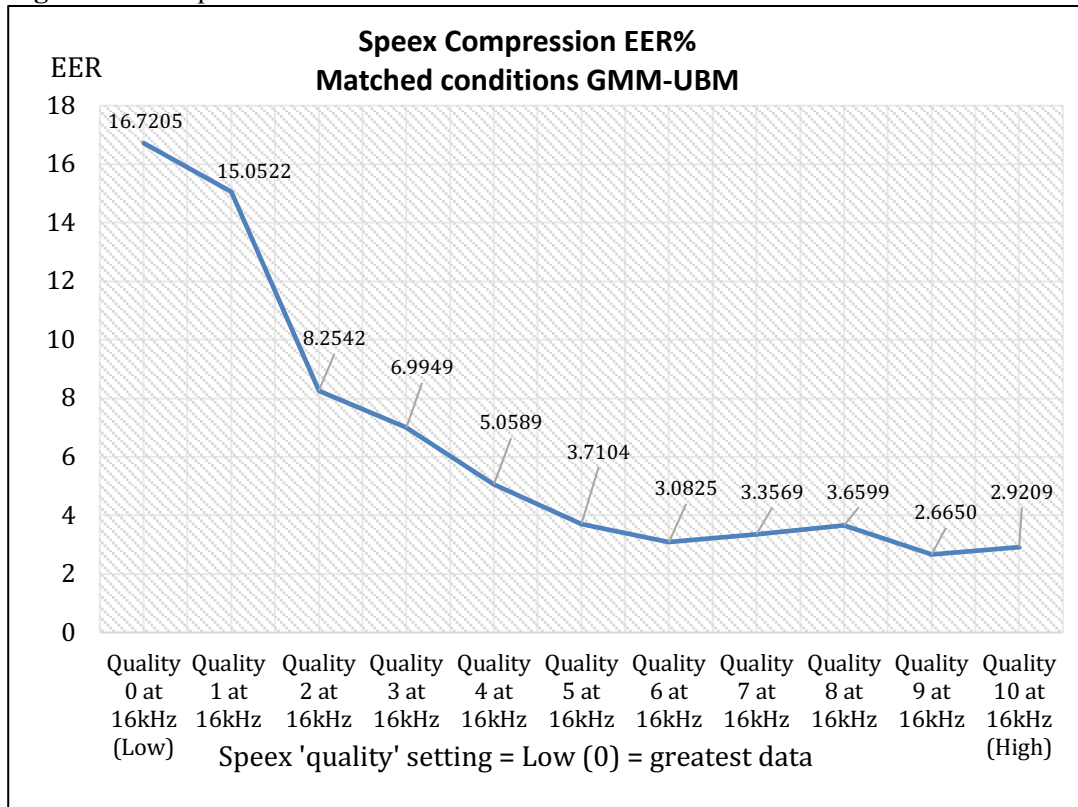


**Figure 11.12:** G.711 EER% results i-vector/PLDA

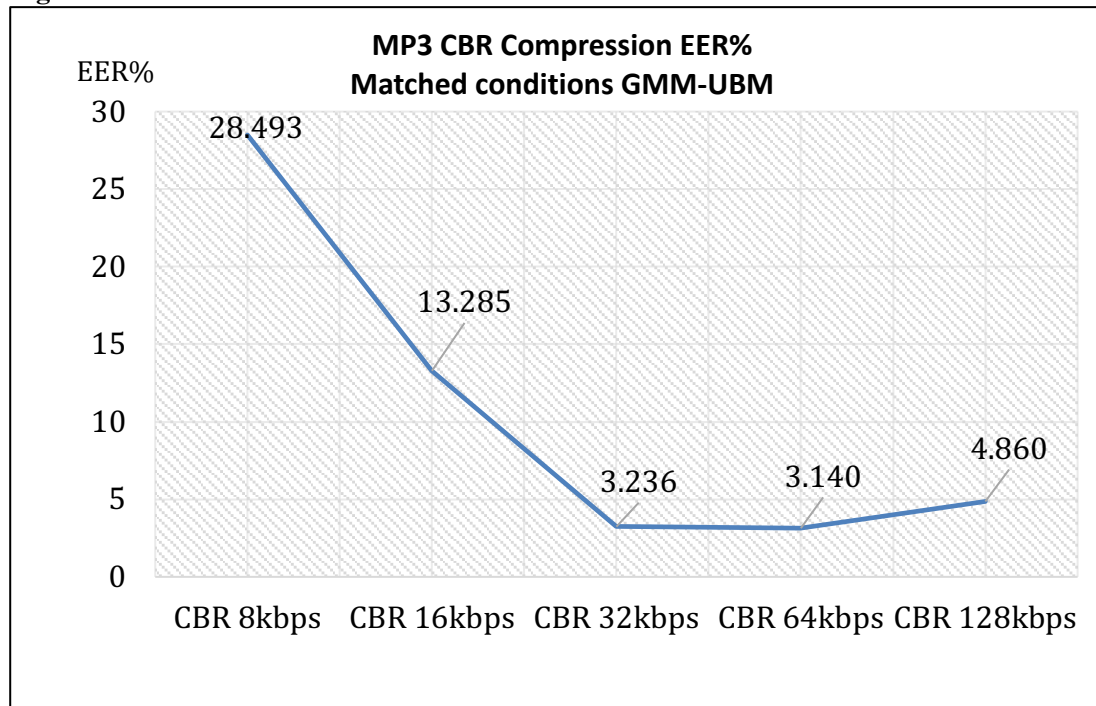


Note lowest EER% (on uLaw) at 6kHz (Figure 11.12), likely due to partial loss of F3.

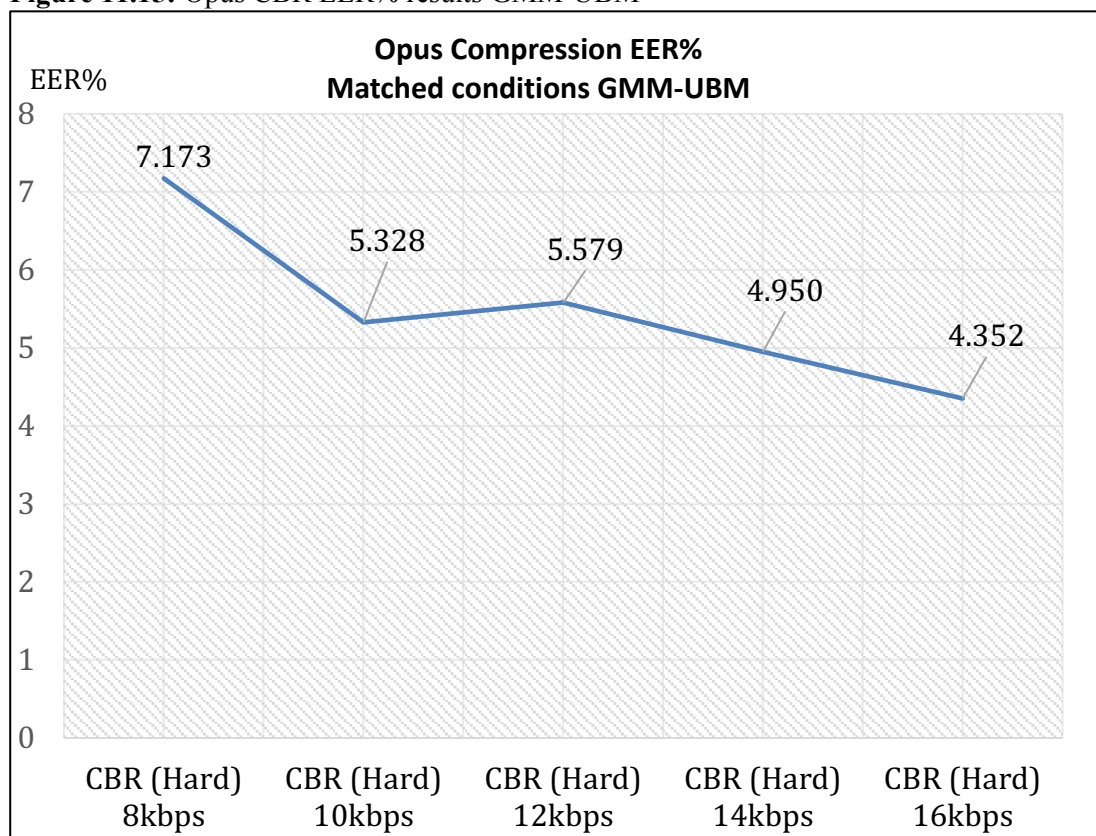
**Figure 11.13:** Speex EER% results GMM-UBM



**Figure 11.14:** MP3 CBR EER% results GMM-UBM

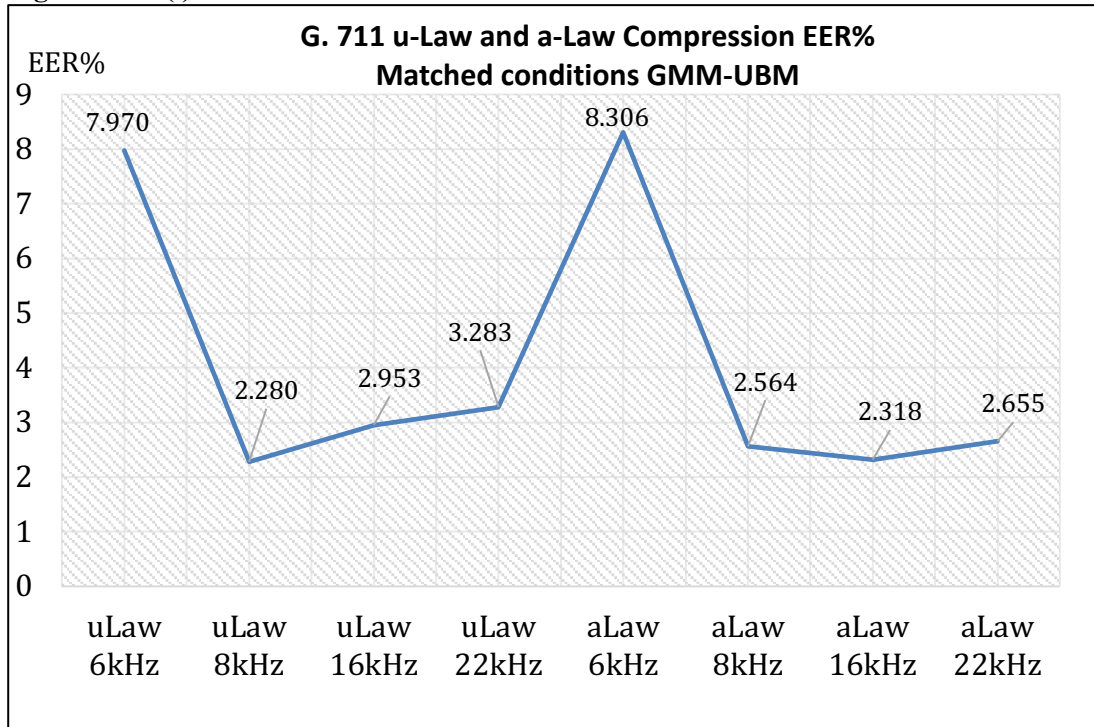


**Figure 11.15:** Opus CBR EER% results GMM-UBM





**Figure 11.16(i): G.711 EER% results GMM-UBM**



Note the two peaks showing poorer EER% performance for aLaw 6kHz and uLaw 6kHz results. This is likely due to the degradation/loss of F3 which is less affected in the aLaw 8kHz and uLaw 8kHz (and higher sample rate) data.

**Figure 11.16(ii): ADPCM EER% results GMM-UBM**

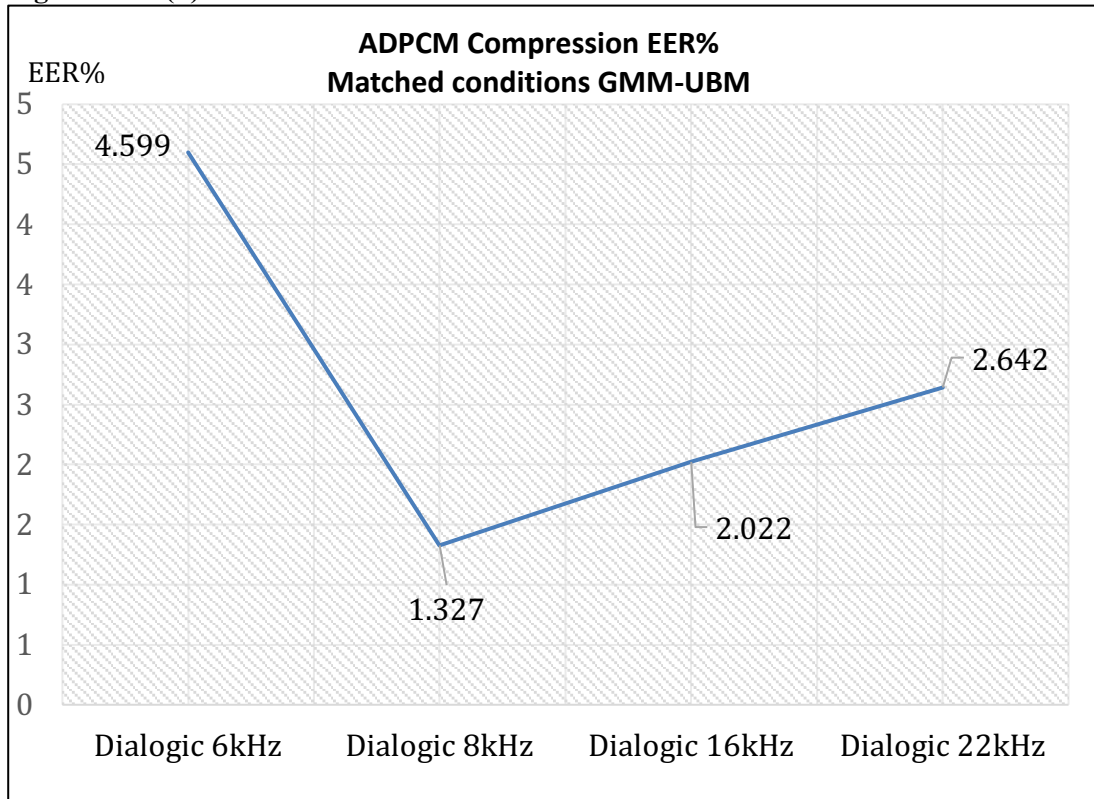


Figure 11.17(i): Baseline zoo plot. GMM-UBM ASR Vocalise (1 of 2)

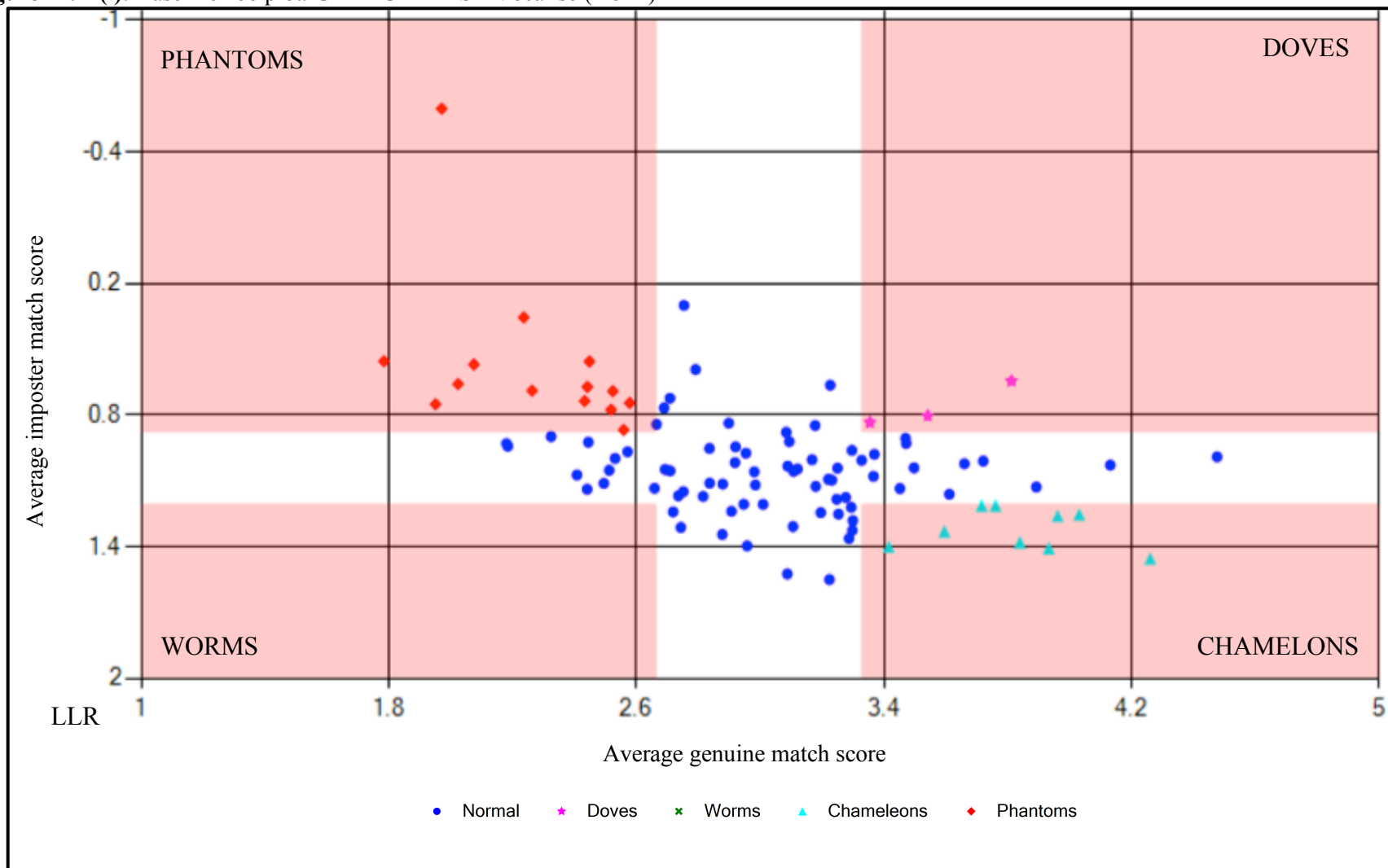
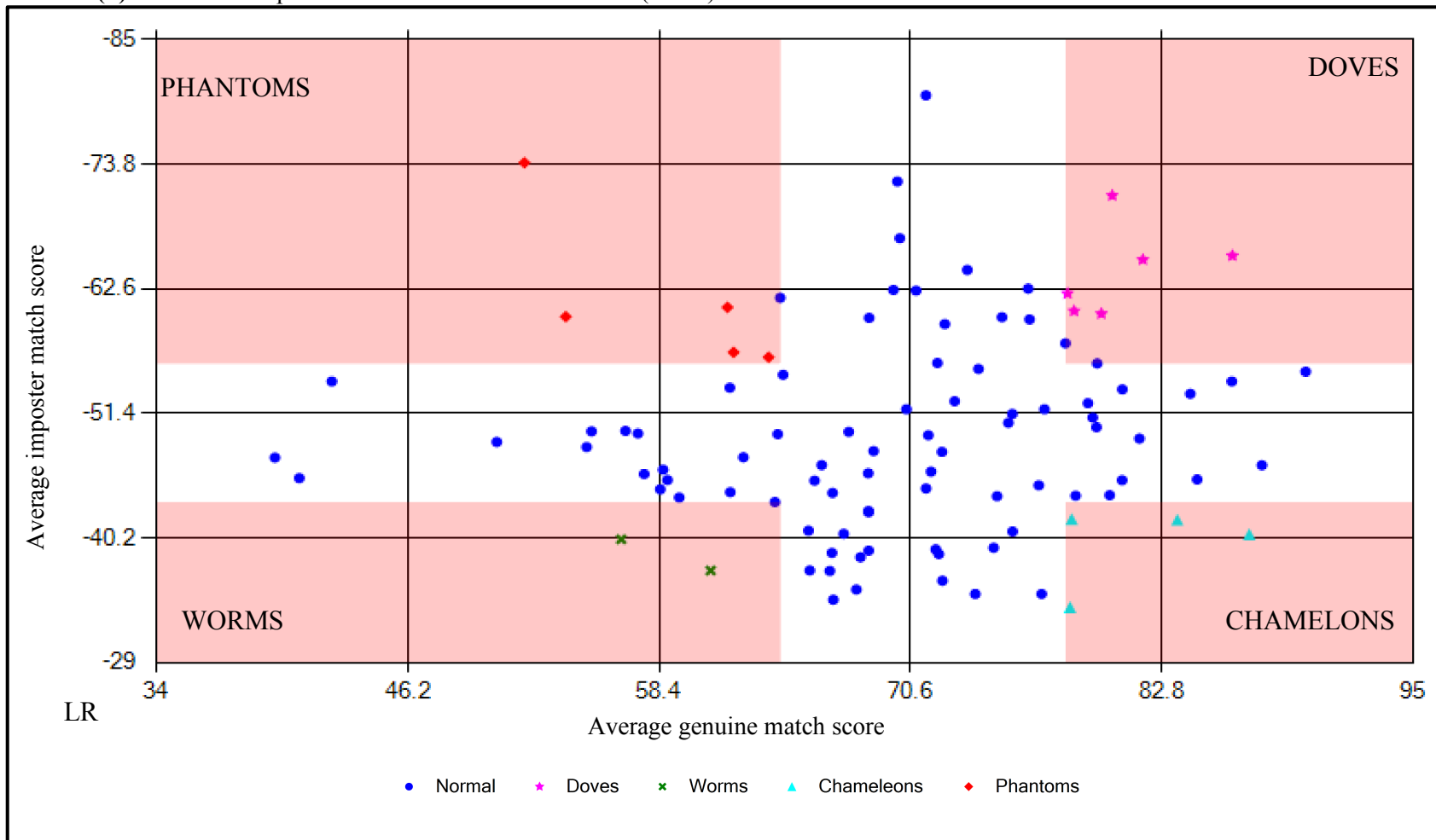
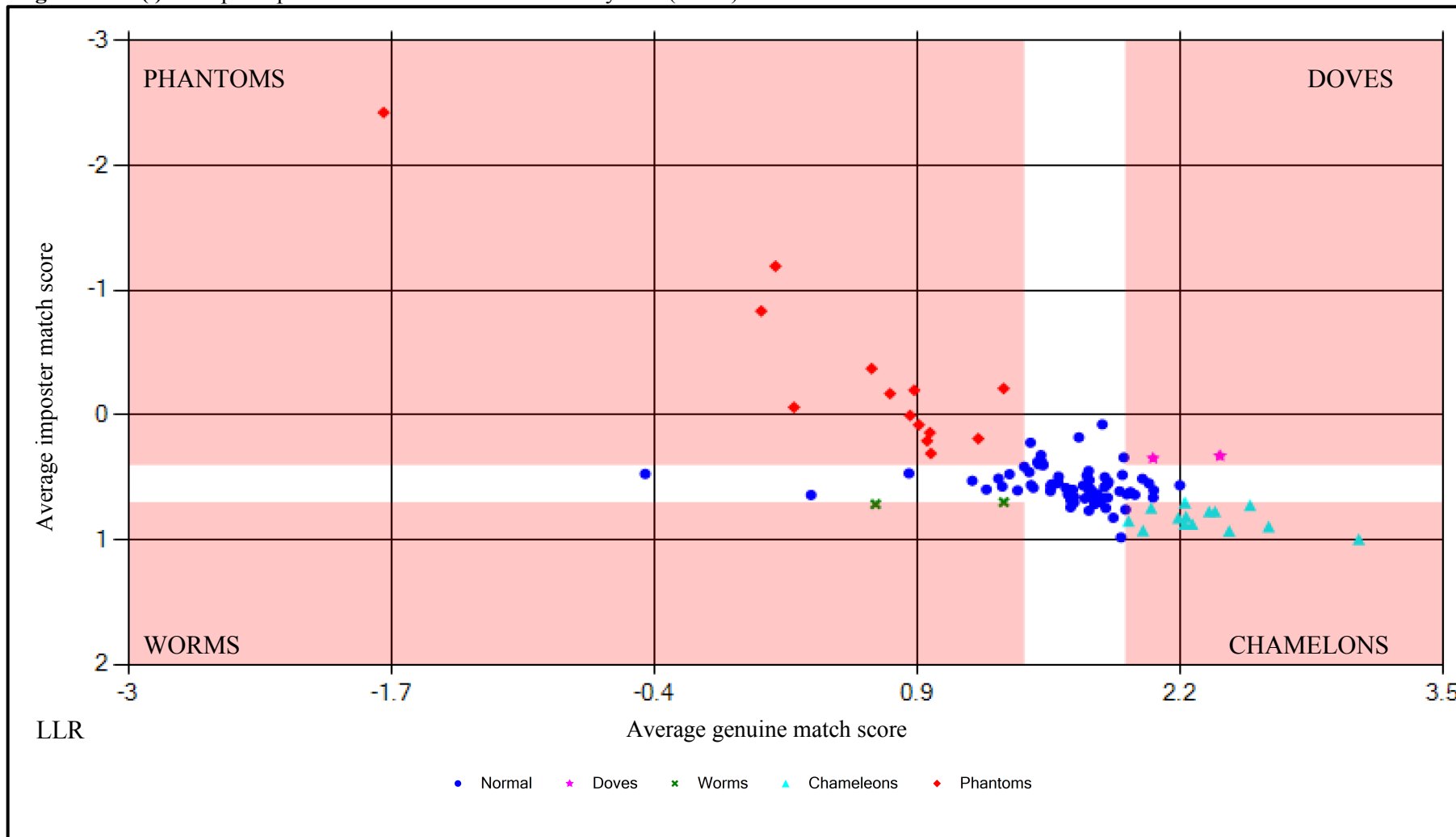


Figure 11.17(ii): Baseline zoo plot. i-vector/PLDA ASR iVocalise (2 of 2)



Note: I-vector/PLDA system produced fewer Phantoms and Chameleons (poor performing speakers) and more Doves (high performing speakers) in comparison to GMM-UBM.

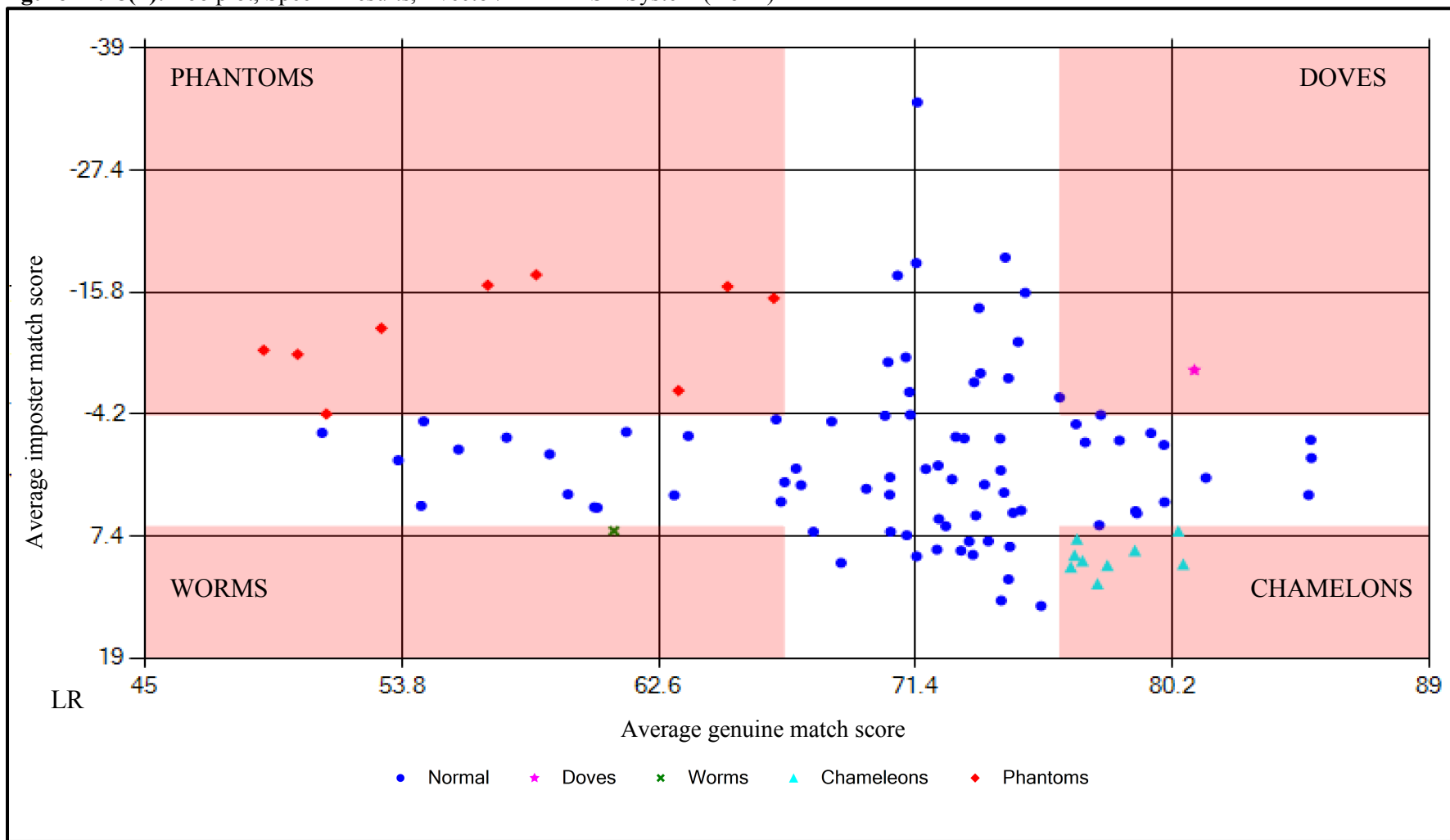
Figure 11.18(i): Zoo plot Speex 1 results GMM-UBM ASR System (1 of 2)



Note: Many more Chameleons and Phantoms (poor performing speakers) with low scores overall, in comparison to baseline.

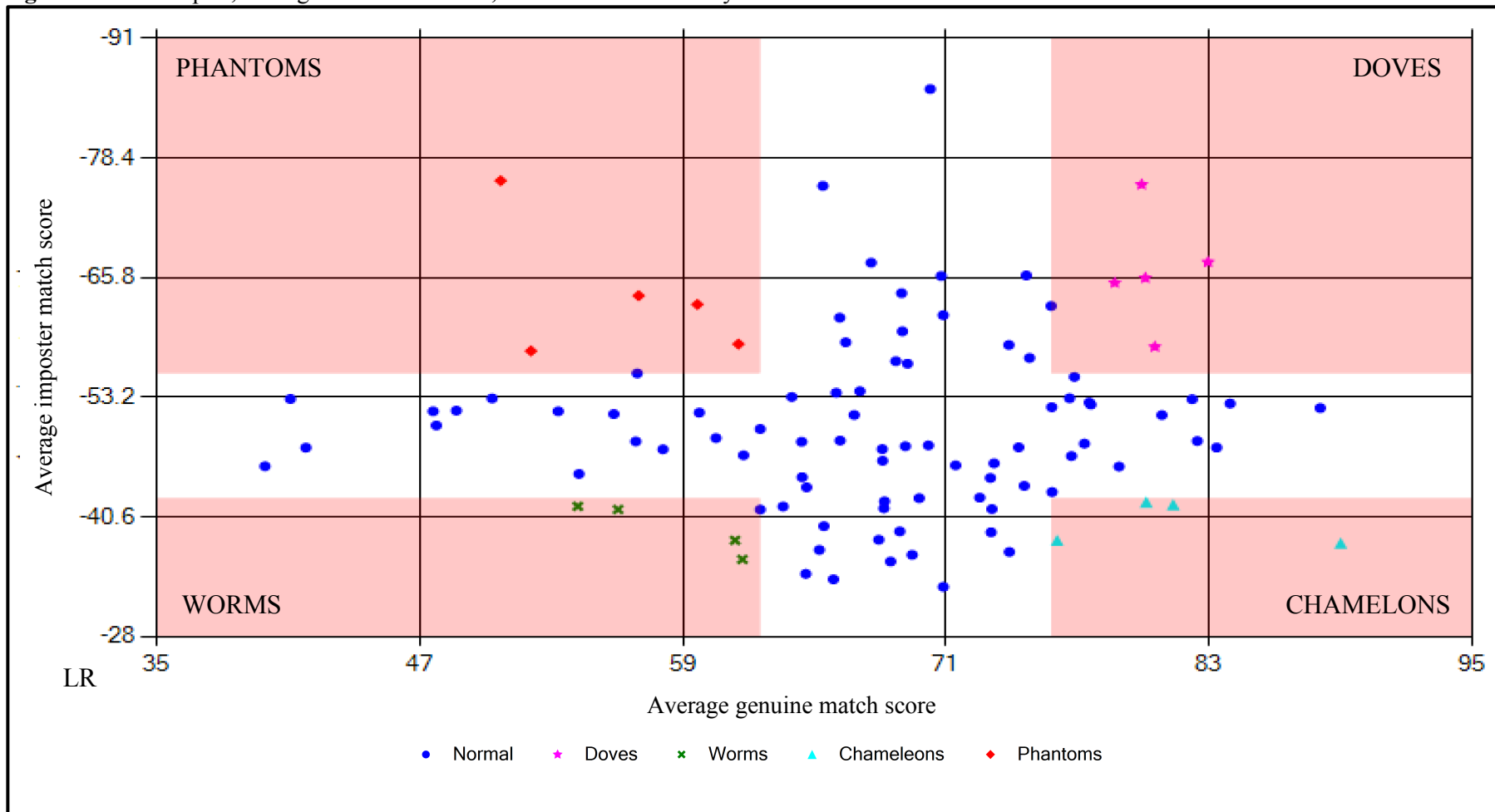
General distribution indicates poor ASR performance overall.

Figure 11.18(ii): Zoo plot, Speex 1 results, i-vector/PLDA ASR System (2 of 2)



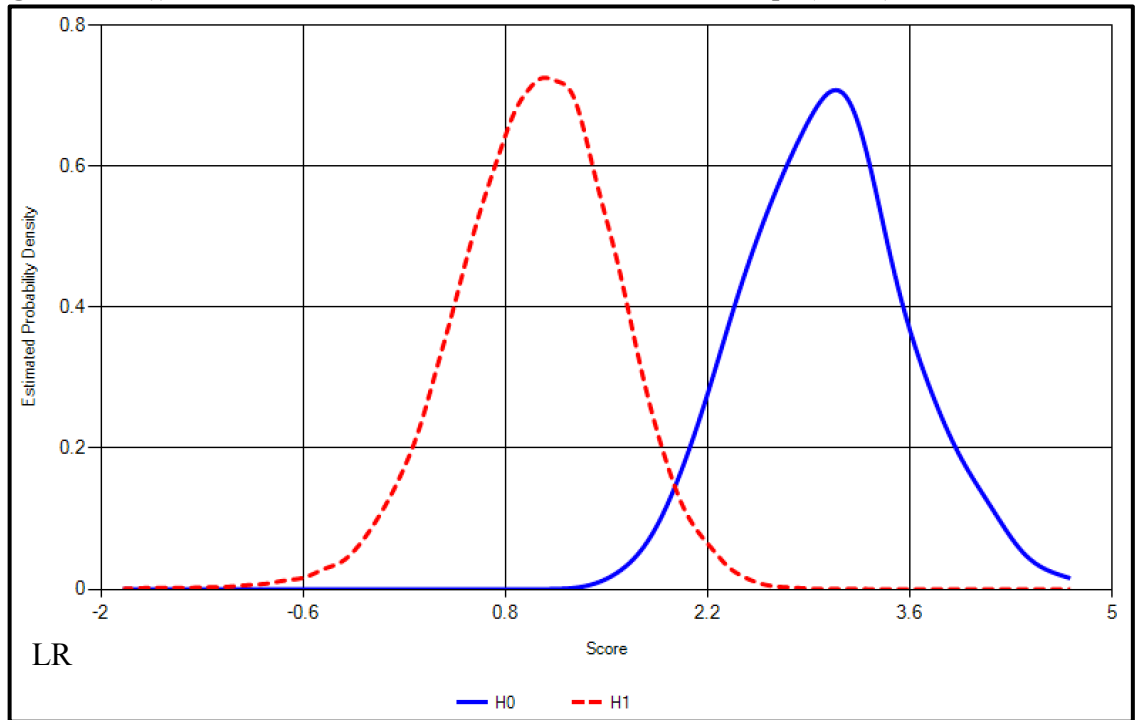
Note: Increase in Phantoms and Chameleons, less Doves, higher imposter match scores, lower genuine match scores (in comparison to baseline).

Figure 11.19: Zoo plot, Dialogic ADPCM 16kHz, i-vector/PLDA ASR System

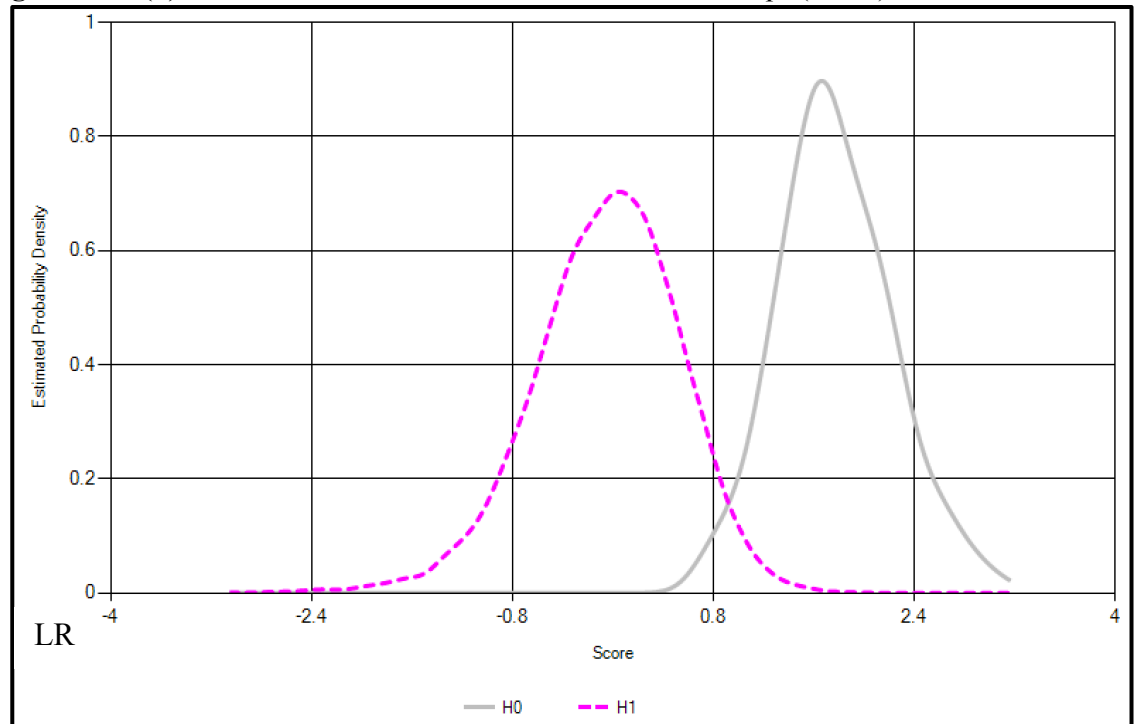


Note marginal EER% performance improvement (+33% relative to baseline) and the increase of poor performing speakers (worms) with movement towards the lower left quartile (in comparison with 11.19(1)). Both indicative of poorer overall performance.

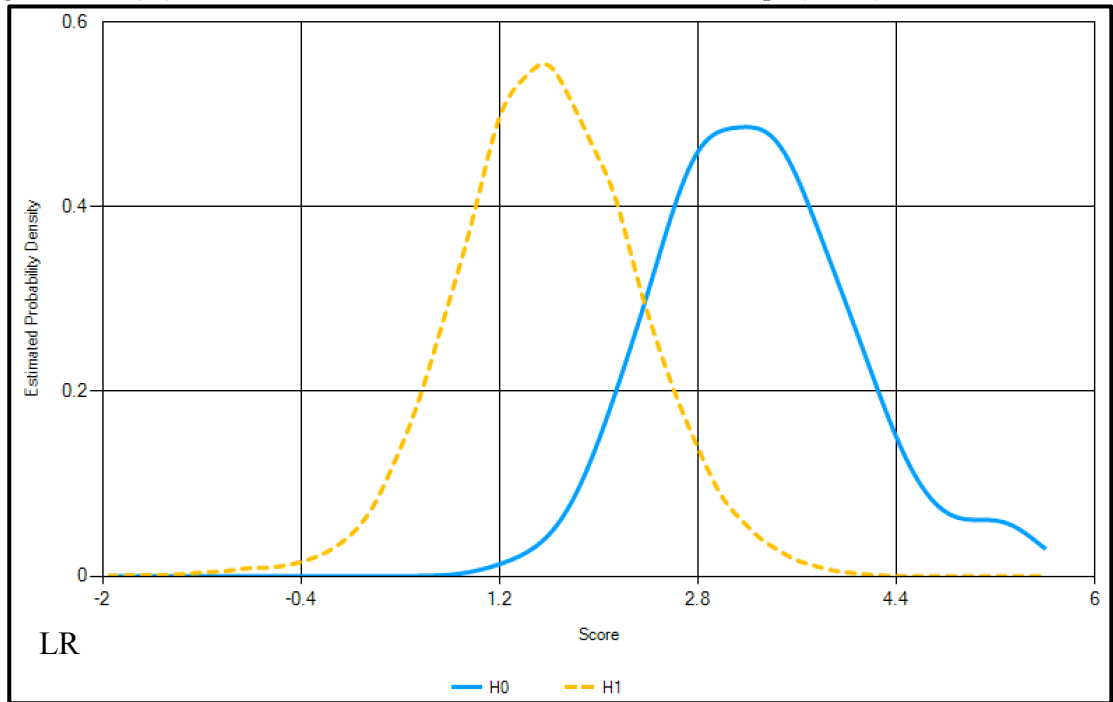
**Figure 11.20(i):** LR Plot. Vocalise GMM-UBM MP3 CBR, 128kbps (1 of 4)



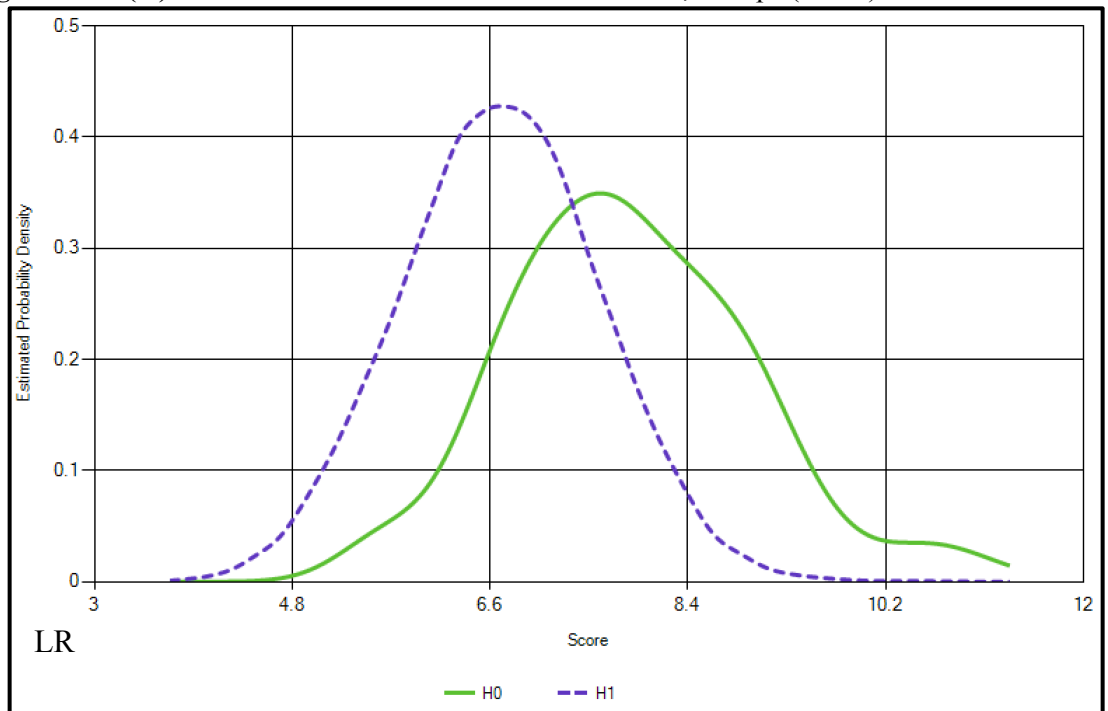
**Figure 11.20(ii):** LR Plot. Vocalise GMM-UBM MP3 CBR, 32kbps (2 of 4)



**Figure 11.20(iii):** LR Plot. Vocalise GMM-UBM MP3 CBR, 16kbps (3 of 4)



**Figure 11.20(iv):** LR Plot. Vocalise GMM-UBM MP3 CBR, 08kbps (4 of 4)



Note the convergence of true positive and true negative distributions as data compression increases, causing potential implications with threshold setting. Also observed was an initial decrease of scores H0 and H1 LR scores and then elevation, particularly at 8kbps (see x-axis legend from Figures 11.20(i) to 11.20(iv)). In discussion, several practitioners also observed this (Alexander, Jessen, Becker, French and Harrison - personal communication and conversation, 2011 to 2013). This is further discussed in 11.8.



## 11.7 Responses to Questions

**Q1 Recap**      **How resilient are more modern i-vector/PLDA ASR systems to codec degradation in comparison with GMM-UBM systems?**

**A1**      As predicted, EER% and Cllr scores improved in the i-vector/PLDA ASR results when compared to those from the GMM-UBM system. In addition, the i-vector system had a better acceptance to ingest for transcoded speech (11 tests produced errors in the GMM-UBM system compared with zero). On consultation with OWR it was suggested that this was likely to small but important improvements made in the speech detection stage for iVocalise.

**Q2 Recap**      **To what extent does ASR performance degrade when transcoding processes are applied to baseline data?**

**A2**      Both systems performed, as expected, exceptionally well on studio quality data. As hypothesised since transcoding tends to remove data through compression and/or band limitation ASR performance degraded in varying degrees. This was proportionate to the extent of the data removed with particular relevance to the speech frequency band (approximately >50Hz to 16kHz) and in line with those bands that pertain more to speaker discrimination (Künzel, 2001; Besacier et al., 2000; Byrne and Foulkes, 2004).

**Q3 Recap**      **How do compression codecs differ in regard to ASR performance?**

**A3**      Whilst each codec type conducts data compression in different ways there were some broad consistencies found (see Q4).

**Q4 Recap**      **Can any operating thresholds be extrapolated relating to data compression rates which may assist with informing ASR use?**

**A4**      As hypothesised, the setting of performance thresholds based on a small set of results from just two ASRs with a significant number of variables is not possible. However, some consistencies in i-vector ASR results were found which could assist with informing or optimising wider system configuration:

- i. MP3 CBR degraded ASR performance below 64kbps and this effect was much more noticeable as compression rates decreased to 8kbps.
- ii. MP3 VBR consistently performed better at quality level 4 (8-16kbps) than quality 9 (lowest) on any setting and performance further decreased as kbps values increased above 64kbps.
- iii. Speex transcoding consistently degraded results at all settings and this was very noticeable below quality 5. Quality 8 performed better than any other setting (EER% and Cllr) including the highest (10).
- iv. All AMR settings produced poor results, though all kbps rates were < or = to 12.2kbps.
- v. All Opus settings below 16kbps degraded ASR performance and the lowest setting (6 kbps) produced a large decrease (i.e. non-linear degradation).

- vi. OGG quality setting 1 produced acceptable results in terms of EER% (0.0051% baseline to 0.0067%) in comparison to quality setting '0', which was shown to degrade performance to a much greater extent (baseline EER 0.0051% to 0.3047%).
- vii. M4a AAC 120kbps had a negligible influence on performance. Data bandwidths of 60kbps produced marginal losses. The lowest setting (12kbps) would not be recommended. However data rates, in relation to EER%, varied dependent on codec and so would not make an ideal acceptability criteria in their own right.
- viii. GSM at standard settings (8kHz) did not significantly negatively influence performance.
- ix. G. 711 and ADPCM are frequency band limiting in nature. Performance was close to baseline at 16kHz with some settings marginally improving performance. Settings of 6kHz and below would not be recommended.

## 11.8 Findings

As predicted, transcoding had a predominantly negative influence on ASR performance. This was consistent for 38 out of 53 x i-vector/PLDA ASR experiments and 46 out of 53 x GMM-UBM ASR (i.e. 79.24% combined).

For the i-vector PLDA ASRs experiments there were 12 instances where transcoding had no discernible effect on EER% and Cllr performance. This performance was not reflected in the GMM-UBM experiments (zero instances).

In line with research from Silovsky, Cerva and Zdansky (2011) several codecs actually produced a small positive effect on ASR performance. There were similar performance gains noted across both i-vector and GMM-UBM systems for Dialogic ADPCM 16kHz and 22kHz (i-vector), 8kHz (GMM-UBM), G. 711 8kHz and 16kHz a-Law and u-Law. The reason for this is not fully understood, but it is suggested that this could be due to the quantisation of digital values when transcoding (i.e. prior to MFCC feature extraction). Effectively the 'rounding up' of digital values could enhance the efficacy of the statistical modelling. Alternatively, or in addition, it could simply be that the normative data (or UBM) is much more densely populated with speech data that has been digitised using those codecs or that the feature extraction method is more effective on the output from certain codecs.

Results showed that score height (LR or LLR output) cannot always be relied on when assessing certain types of transcoded data, particularly very low bit rate perceptual codecs (MP3, M4a). High true negative scores were evident and this was more prevalent for the GMM-UBM system. This should be factored into analysis. Higher H1 scores were noted for transcoded data and this could

also impede the setting of thresholds – it would also suggest against combining certain types of transcoded data alongside non transcoded data.

Codecs can have other options available within their settings. For example; speech detection, gain control and the addition of comfort noise for filling gaps in conversational speech. In summary, having an awareness of the specific technical influence for different types codecs and, ideally, the transcoding history will be important in determining whether ASR comparison is practical and can be conducted within acceptable performance boundaries. It can also be helpful when interpreting ASR results and should be incorporated into reporting.

Codecs set to their highest, most severe compression rates (e.g. MP3 8kbps, Speex quality 1-4, AMR, Opus and M4a) had the most negative effect on EER% relative to baseline. The difference in types of codec proved important with the perceptual codecs (MP3 and M4a) performing worse than all other codecs at their highest compression settings. One explanation for this could be because perceptual codecs are generally optimised for ‘human’ listening (e.g. music) rather than specifically speech and machine analysis. Perceptual codecs remove data from the acoustic signal that is not always perceivable (by humans) but may still be useful to the ASR system for discrimination, for example. The acoustic signal is scanned much more comprehensively by the machine, effectively at all frequencies within the bandwidth set by the operator - noting that MFCC’s apply the Mel scale (which generally pertains to human hearing). It is suggested that the effect of perceptual codecs would also worsen with the addition of noise prior to transcoding as the codec may effectively ‘prioritise’ noise over speech.

At lower compression settings (i.e. higher quality), the psychoacoustic codecs produced no discernible degradation of performance and EER% matched baseline. Opus, optimised for speech, performed surprising well at 16kbps producing an EER% of 0.0051 identical to baseline (i-vector/PLDA). Poor performance was evident in respect to EER% relative to baseline and artificially raised mean H1 values were noted for the i-vector/PLDA ASR results (true negative outcomes). A combination of elevated H1 and H0 values were also shown in the GMM-UBM results. The raising of score distributions for GMM-UBM results was particularly apparent for experiment 13 (MP3 CBR 8kbps) where the H0 and H1 means raised from 3.017 and 1.008 for baseline to 7.782 and 6.693. This could clearly have implications for casework if not compensated for.

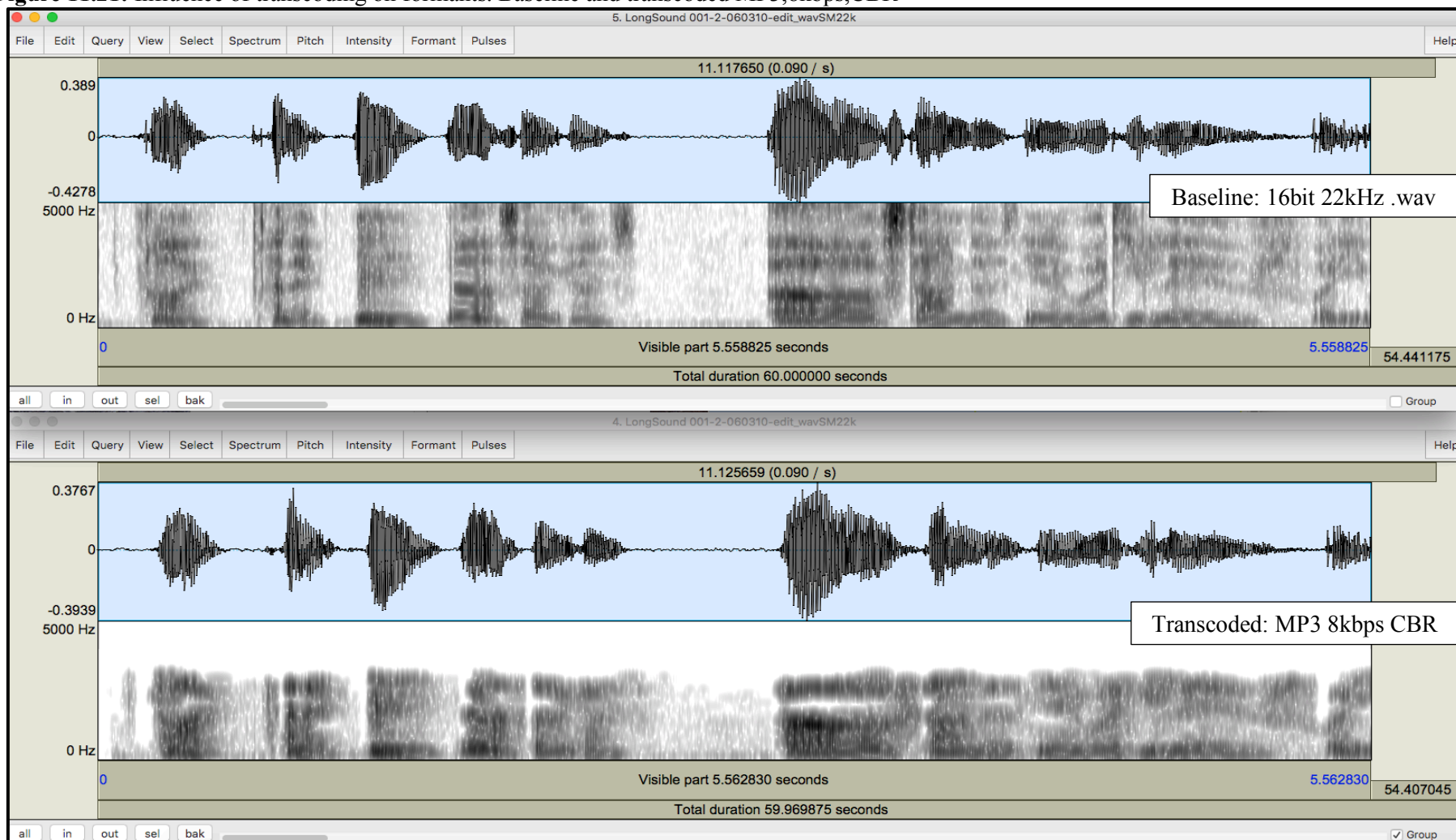
The i-vector/PLDA system outperformed the GMM-UBM ASR in all instances with regard to EER% which was, overall, poorer throughout. All experiments were successfully processed for the i-vector/PLDA ASR. For nine GMM-UBM ASR experiments, the quantity of files passing the VAD file-ingest stage fell. Statistics were therefore generated on successful file comparisons – although this effectively distorts the success rates for the GMM-UBM ASR. Note that there were two

experiments for which all files were rejected and these were logged as system errors (GMM-UBM only) – this has practical implications for the technical acceptance of heavily compressed files for ASR comparison.

## **11.9 Additional Tone Experiment**

Throughout the transcoding batch processed audio files were checked for technical quality. During this process, files were examined in Praat using spectrograms. As expected, the effect of data compression was often both visible and audible. Watery artefacts often referred to as ‘chiming’ could be heard and upper speech frequencies were often heavily muted or inaudible. These auditory effects would certainly appear prominent to a forensic practitioner. These effects were often evident in the spectrogram too with the removal of energy at points in the frequency range, in the example below, removing F4 through frequency bandwidth reduction. Figure 11.21 shows identical utterances from Speaker 001 transcoded using MP3, CBR, 8kbps.

**Figure 11.21:** Influence of transcoding on formants. Baseline and transcoded MP3,8kbps,CBR

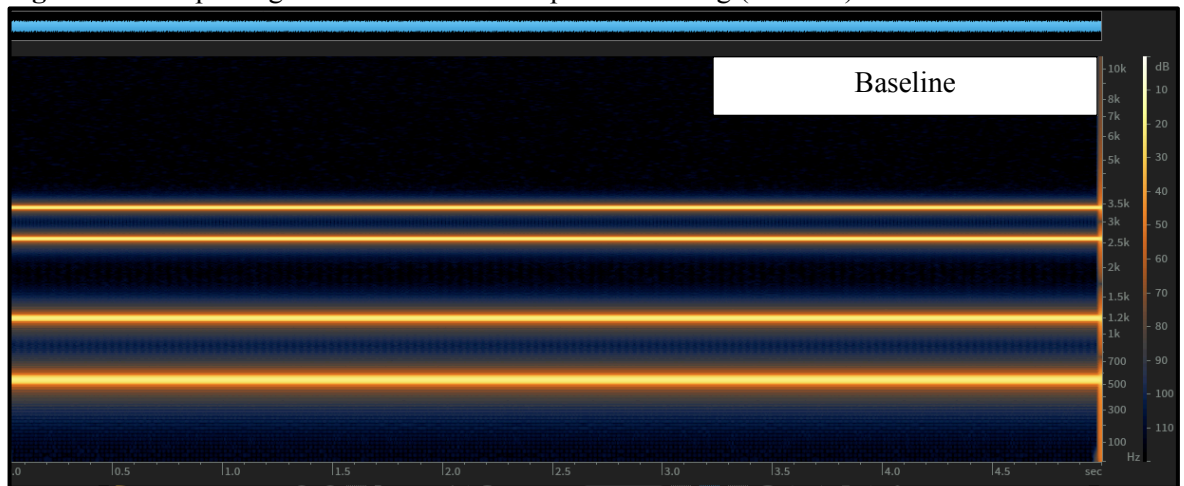


Note that the examination of the waveform (amplitude) would yield only marginal between the two files in comparison to the spectrogram. Transcoding, to this extent, would present issues both to a forensic practitioner and ASR analysis - producing data loss and constraining frequency bandwidth.

It was noticed that the formants in transcoded speech also appeared to be very marginally shifted in the frequency domain in comparison with baseline (.wav) tone values. This required further examination.

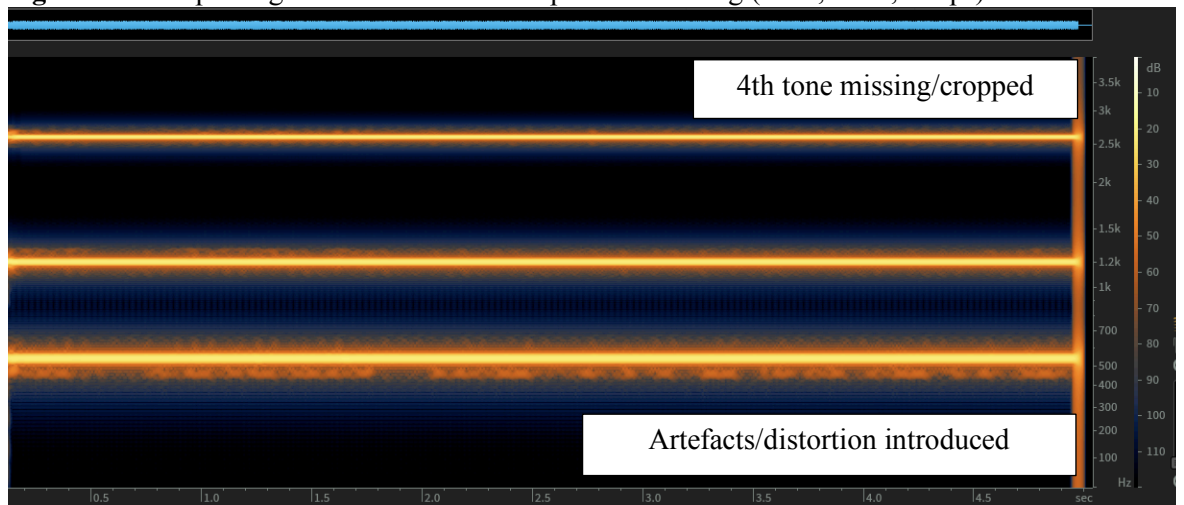
A brief experiment was generated using only a set of constant test tones which very broadly simulated several key (mean) frequency points of speech (550Hz, 1,200Hz, 2,600Hz and 3,400Hz). Note that 4 distinct tones were generated, as opposed to a single tone with multiple harmonics. The motivation for this method was to more closely examine the frequency shifting whilst removing the variability within the speech formant data, to better quantify this effect. The 4 test tones were generated in iZotope RX Advanced (v.6) and were transcoded using identical settings from the main experiments. The two codecs examined were MP3 (set to CBR, at 8kbps) and Speex (quality setting 8). Mean frequency values were estimated using Praat at 16 interval points over 1s.

**Figure 11.22:** Spectrogram view of test tones pre transcoding (baseline)

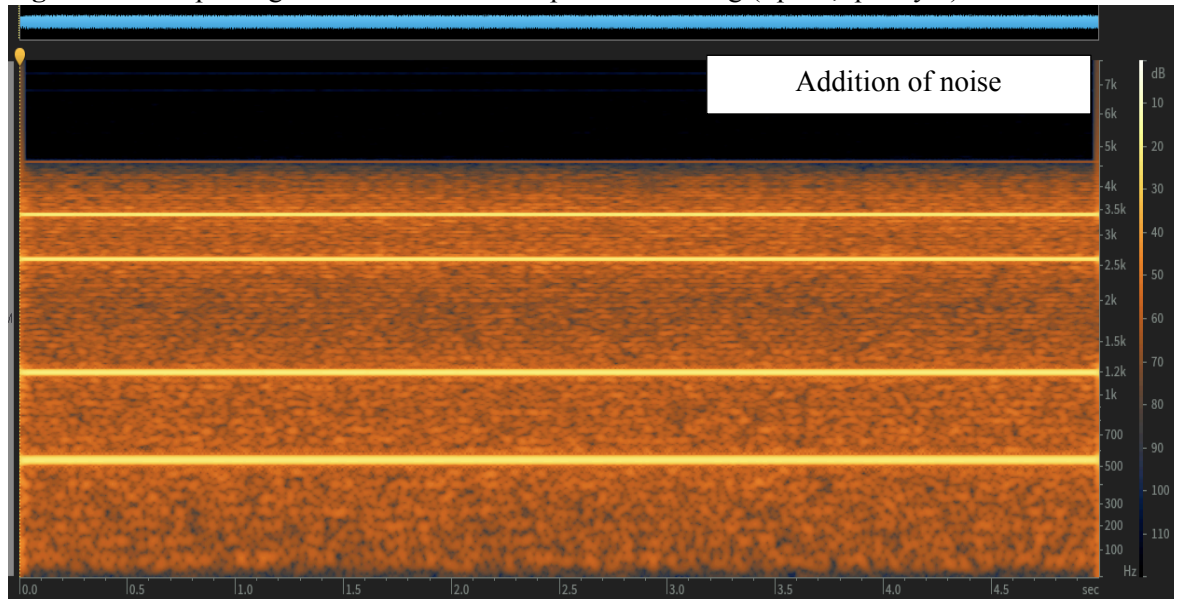


Please note the difference in y-axis scales between Figures 11.22, 11.23 and 11.24.

**Figure 11.23:** Spectrogram view of test tones post transcoding (MP3, CBR, 8kbps)



**Figure 11.24:** Spectrogram view of test tones post transcoding (Speex, quality 8)



Measurements of the test tone frequencies were extracted using Praat and these are summarised in the results Table 11.25 (full data in Appendix J).

**Table 11.25:** Mean frequencies, 3 additional transcoding experiments, 4 test tones

Codec	Test tone 1 (mean)	Test tone 2 (mean)	Test tone 3 (mean)	Test tone 4 (mean)
<b>Baseline</b>	<b>550Hz</b>	<b>1,200Hz</b>	<b>2,600Hz</b>	<b>3,400Hz</b>
MP3, CBR, 8kbps	540Hz	1,200Hz	2,580Hz	N/A
Speex, setting 8	563Hz	1,206Hz	2,600Hz	3,400Hz

**Green** = 0 variation in reference tone

Measuring formants accurately can be problematic (Harrison, 2013) however the lack of variation measured for several tones (MP3, tone 2, Speex tones 3 and 4) suggests that the measuring process itself is effective. However, clearly the dataset is very small and results should be treated with much caution. Although the differences, post transcoding, are very marginal (tone 1, Speex tone 2 and MP3 test tone 3) it appears that transcoding is influencing the accuracy of the tone measurements and this appears more noticeable for lower frequencies. It is also possible that acoustic distortion could be influencing the mean estimation – due to the addition of artefacts, which appear to broaden the frequency bandwidth of the lower tone in the MP3 example (Figure 11.23).

Pure tone and speech are quite different in acoustic complexity and the effect is not likely to directly transition (from tones to speech). Nonetheless, it is possible that altering data in the frequency domain, could confuse the automated feature extraction or measurement process with respect to speech information (e.g. formants). This could help to partially explain why, in the experiments completed, the ideal/matched conditions generally performed better, since the acoustic changes

would manifest as similar for both SM and TA although this explanation does not hold if F3 is severely degraded/missing. The distortion of the frequency domain, from different types of transcoding, requires further research – but there are likely similarities with mobile phone transmission and the influence found on formant frequencies F1 and F2 (Künzel, 2001; Byrne and Foulkes, 2004; Jessen and Becker, 2010).

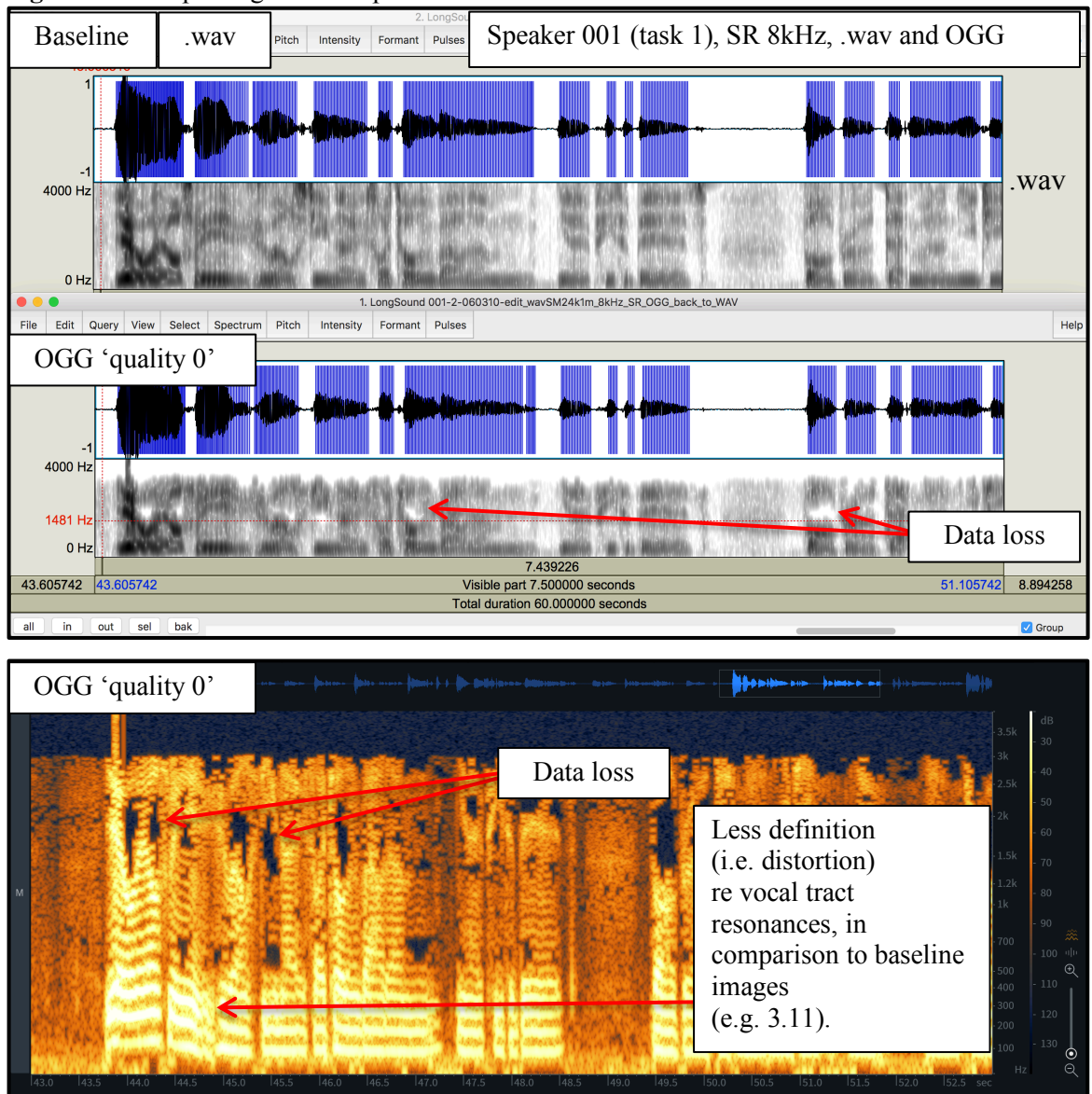
With the increase in the application of ASR systems (Morrison et al., 2016; Gold and French, 2011; 2019) in addition to the diversification in communication methods (e.g. VoIP) it is expected that we will continue to see a greater application of ASR's on transcoded speech. It is hoped that the comprehensive new data provided in this chapter contributes to the field through providing metrics and elevating awareness of the influence of transcoding on ASR system performance.

## **11.10 Discussion**

There are many variables relating to codecs which have been shown to influence the performance of ASR systems on transcoded speech. Whilst very low bit rates (kbps) tended to degrade performance to a greater extent, bit rate cannot be used as a measure in itself to solely determine audio acceptance for ASR analysis. The type of codec and compression settings (CBR, VBR, Quality) are clearly relevant too and determine the parts of the frequency spectrum that are affected (compressed or removed). Codecs that limit the frequency bandwidth or prevent speech passing the speech detection phase (i.e. also inhibit) degraded performance and data loss was visible in spectrogram analysis (Figure 11.26).



**Figure 11.26:** Spectrogram examples of codec distortion and data loss.



The data loss is predominantly in the mid to upper frequency range where discriminative speech content is present, influencing attribution. Vocal tract resonances also appeared blurred (Figure 11.26, lower spectrogram) suggesting codec distortion – which would, in turn, degrade the accuracy of feature extraction and increase speaker confusability.

The later generation i-vector ASR outperformed the GMM-UBM system with notably better EER% and Cllr. It was also more acceptant of degraded speech material, this is likely due to more accurate speech detection, improvements to the feature extraction process and the much larger normative dataset (PLDA). As a practical recommendation ASRs should be upgraded regularly to ensure they are benefiting from continual advancements that are made. This, however, may present problems with regards to replication of results and moves to provide more universal normative sets.

Finally, psychoacoustic codecs set to low bit rates were shown to produce variable outcomes for score height (Figure 11.7, GMM-UBM system) including artificially elevated scores. At worst, this

could cause the incorrect interpretation of a high score as a verification. As also seen in Figure 11.20 transcoding can cause the convergence of score distributions for true positive and true negative outcomes, making threshold setting more problematic, particularly if a mixture of transcoded and non-transcoded was compared.

# Chapter 12 Discussion

---

This chapter discusses all of the experiments completed and provides additional explanation for the results obtained, placing them in the context of the third objective for this thesis. To recap, the third objective was to provide recommendations with regard to ASR suitability for forensic speaker comparison (FSC) application and investigative use based on the research and experiments completed and within the context of acoustic variability.

## 12.1 Summary of Automatic Speaker Recognition Performance

Overall, results showed that in almost all instances contaminants and inhibitors had a negative influence on ASR system performance (EER% and Cllr). It was demonstrated that when acoustic degradation was most severe, ASR performance was reduced to almost chance equivalence (i.e. system fail). Conversely, error rates were impressively low when degradation was light to moderate.

The i-vector ASR system proved to be much more resilient to acoustic variability and, in almost all instances, outperformed the GMM-UBM system – often by a wide margin. In summary, this is likely due to the additional detail inherent in the statistical modelling and comparison phases, generational advancements (e.g. small improvements to the feature extraction and VAD phases) and the use of much larger normative datasets which provides additional improvements in statistical comparison. The results above are broadly consistent with similar research studies completed both before, during and after the experiments completed in this thesis.

### 12.1.1 Tipping Points and Acceptability Criteria

Performance tipping points were identified when acoustic degradation bordered from moderate to severe and these were observed in both types of ASR systems. Whilst occurring at slightly different points for the GMM-UBM and i-vector system, performance consistencies were found within the 5 conditions. In summary, tipping points were observed when:

- i. Total net duration was less than 10s for SM or TA, or the total net duration of both SM and TA was very unevenly distributed (e.g. 5s for SM and 55s for SM);
- ii. Noise was introduced at -20db to -15db. This threshold will also be influenced at a speaker level by vocal effort/initial SNR – although pre-emphasis (MFCC) is likely to compensate to some extent;
- iii. Reverberation was added that was >1s (T60) and/or the reverberant reflections were very complex;

- iv. Frequency bandwidth limiting (from filtering and/or transcoding) effectively deleted speech data below 3.5kHz (i.e. removal of areas of the speech spectrum which provide speaker discrimination information);
- v. Data compression rates degraded frequency bandwidth, as per point (iv), and/or loss of speech data occurred which limited net duration and/or kbps fell below a certain level. Importantly, data rate (kbps) in and of itself was not a tipping point and should not serve as an acceptability criteria without consideration, and in reference to, the specific codec type and settings (11.6).

Note that tipping points will vary depending on ASR system, normative data, settings and data quality (SM, TA and normative).

In reviewing the entirety of the experiments completed, the interconnectivity between the different forms of contaminants and inhibitors also became more apparent. Transcoding often limited frequency bandwidth and net duration. Reverberation and SNR were shown to limit net duration (i.e. VAD). Reverberation and transcoding influenced and/or constrained frequency content.

Finally, intrinsic speaker variability (e.g. vocal effort, voice quality and pitch) must also be factored into acoustic variability since they are, in many cases, linked variables and are effectively the starting point from which degradation can then occur.

Zoo plot positioning for speakers demonstrated that speaker performance did not degrade in a uniform way across all speakers with regard to acoustic variability. An example of this was shown where speakers with quieter voices were more likely to yield lower SNR recordings than high SNR and in turn perform poorer when compared on ASR systems. Comparing low SNR recordings with high SNR recordings (i.e. unmatched SM and TA) was shown to be problematic and not advised. Joining recordings together of highly divergent acoustic quality (for either TA or SM) was not within the scope of the experiments but would certainly not be advised either, based on results. Extracting WADA SNR values for all speaker models on enrolment would be recommended with respect to identifying those speakers of poor SNR and assist with managing line ups/watch lists accordingly.

## **12.1.2 Automatic Speaker Recognition Performance Metrics**

Experiments have demonstrated the benefits of fully establishing the performance limits of an ASR system in relation to acoustic variability. Performance limits should be quantified to inform audio acceptability criteria during technical assessment. Metrics pertaining to audio quality, e.g. a full technical report and documentation of any process(es) applied, should be included in reporting. A

baseline corpus specifically designed to assess multi-laboratory ASR performance should (consistently) be used in this regard and a new ‘forensic\_eval\_01 dataset’ (Morrison and Enzinger, 2019) offers an opportunity to measure ASR performance, using a standard corpus created for the purpose.

### **12.1.3 Opportunities for Automatic Speaker Recognition Improvements**

Whilst not a specific objective of this thesis, the experiments support the following recommendations in terms of improving ASR performance. In summary, these were:

- i. ‘High Definition’ speaker recognition. Increasing the frequency bandwidth from 0-4kHz, 8kHz SR to 0-11kHz, 22kHz SR (in conjunction with raising the MFCC/feature extraction bandwidth);
- ii. Improving the matching of conditions (SM, TA);
- iii. Adapting the normative dataset to better reflect the channel conditions of the SM and TA;
- iv. Completing audio enhancement/noise reduction;
- v. Adapting ASR settings to:
  - a. Amend the feature extraction settings (MFCC/optimising numbers of features and filters, LTFD/ formants F1, F2, F3, F4);
  - b. Amending and improving the SD/VAD phase – i.e. speech/non-speech discrimination
  - c. Increasing/optimising the numbers of Gaussians - see preliminary tests and also Janicki (2012).

With respect to point (iv) more research is required as results are not consistent regarding the type and complexity of noise. Both ASR performance improvements and degradation can occur, as also found in Künzel and Alexander (2014). In addition, where speech data has been removed by the codec (data compression) it cannot be replaced.

There were several instances where transcoding all the SM and TA data marginally improved ASR performance on both GMM-UBM and i-vector systems (11.6). As stated in (iii) a plausible explanation for this is that it is likely due to improved relevance of normative data (i.e. more normative material transcoded with the same codec as the SM/TA). However, as the full examination of all normative data used was not possible to confirm this, further research is required.

### **12.1.4 Technical Quality Assessments**

Results from the experiments support the inspection of the technical quality of recordings presented for casework comparison. This is consistent with statements from Harrison and French (2010) and

Becker et al. (2012) where recommendations are made that full technical assessments are undertaken as a pre-analysis process for both auditory and/or any acoustic measurements. As stated in Becker et al. (2012: p.5):

‘Using automatic forensic voice comparison systems without any further investigation of the recording material results in a considerable proportion of errors. This proportion can be reduced if forensic phonetic experts are involved to judge the material as well as speaker features’.

Throughout the experiments completed, spectrogram analysis proved essential for examining the extent of audio degradation – although it should be noted that minor degradation, which marginally influenced ASR performance, was not always immediately apparent. Also, technical analysis can be more difficult in the context of casework in comparison to research (i.e. more complex). Nevertheless, spectrogram analysis should form a part of the workflow when using ASR systems. It is also recommended that broader technical quality assessments form an integral part of managing a networked ASR system (e.g. checking quality of end to end network data, particularly with respect to transcoding).

### **12.1.5 Mismatched Conditions**

Channel mismatch is commonly noted as an important factor in ASR performance. In recent research, Hughes et al. (2018) completed examination of DyViS performance using a bespoke LTFD ASR system and noted that mismatch between high quality and low quality (GSM) channel domains was detrimental to overall performance with a difference of 21.66% EER and Cllr 0.46. The group also stated the importance of examining individual speaker performance within systems and recommended the application of zoo plots to examine variation in speaker performance between conditions (2018: p.4).

Results from experiments indicate that great care is required when comparing files of dissimilar audio conditions within the same set of comparisons and confidence of ASR outcome should be adjusted accordingly. Casework examples of this could pertain to supporting ASR imposter line ups with (as close as is possible) uniform conditions – important since, for instance, transcoding using different codecs was shown to artificially inflate or lower H0, H1 and LR score output. It could be argued that score adjustment or completing a control degradation (of the higher quality recording to that of the lower) can mitigate for mismatched conditions. However, in practice, difficult questions arise as to exactly what to calibrate to, if SM and TA are highly divergent, with respect to channel. There may also be a lack of normative data which reflects conditions or it may not be possible to degrade sample(s) in a controlled way without knowing full file provenance/origin of the file(s). Finally, if the technical quality difference between SM and TA is highly divergent (i.e. non-ideal conditions)

then careful consideration should be given before accepting comparison tasking using ASRs at all, particularly if applying an older GMM-UBM ASR system (or earlier).

### **12.1.6 Population/Normative Data**

Much effort was placed during the data preparation phase to ensure high quality speaker models, test audio and normative sets were constructed. This was completed through inspection and the redaction of very poor quality speech data (at an editing level for SM and TA) in addition to care taken in data selection for normative sets. Whilst this had a high resource cost, very low EER and Cllr scores were therefore obtained for the baseline results. The importance of audio quality, particularly relating to the normative set, was studied by Biswas, Rohdin and Shinoda (2015) who stated:

‘The training data of a PLDA model is often collected from a large, diverse population. However, including irrelevant or noisy training data may deteriorate the verification performance.’

Biswas, Rohdin and Shinoda (2015: p.32).

As previously discussed (3.4.7) other research advises to the contrary and the PLDA session experiments completed (9.5.3) suggest that a balance is probably required. This should reflect both the (degraded) conditions (of the SM and/or TA) and maintain a high quality data set overall (total variability) which accurately captures the intricate speech energy patterns and diversity of a large population set. With the improvements shown in i-vector/PLDA advancements, statistical modelling and the large increase in the size of normative/population data, there are opportunities available to further improve them in the context of acoustic variability – and therefore ASR performance as exemplified in 9.5.3. The application of different, and potentially selectable normative sets, reflecting different acoustic conditions, could maintain ASR performance under light to moderately degraded conditions. Further research is required in the area of population data and the data modelling of conditions (see 13.2).

Finally, it is recommended that the population data should either be completely accessible (and configurable) by a fully trained expert, or a comprehensive report provided by the manufacturer as to the technical quality and constitution of the data - i.e. what is in the normative data, what is the channel/domain and acoustic quality of that data (frequency bandwidth, SNR, reverberation times, net duration and codec) and details of the speakers contained therein (i.e. additional information such as language spoken, age, accent, gender etc.). The purpose of this would simply be to assist the analyst with determining/ensuring relevance to the comparison and maintaining it throughout different casework applications.

### **12.1.7 Operator Training and Standards**

With the near infinite variability in speech, ISO standards have not yet been specifically applied to ASR system use. Decisions made by the operator such as editing or adjusting ASR settings can influence outcome and this also provides additional variability. When assessing technical quality, decisions are also required which require operator training in acoustics and benefit from ongoing development (e.g through experience). Nevertheless, there are guidelines pertaining to best practice in the broader context of forensics (ISO17025) and recommendations have been published by both ENFSI (2015), and more recently the UK Forensic Science Regulators (2017), that pertain to speech analysis and process management.

There have also been studies to ensure that the LR framework itself is robust, validated and meets ISO17025, although in the context of other biometric comparison – see Meuwly, Ramos and Haraksim (2017: p.83). Improving standards requires investment in people, infrastructure, training, accreditation, peer review/inspections, research and greater audit/control processes.

The application of better standards to recording process(es), whilst difficult to implement, would also assist with providing output which is more suitable for ASR comparison. Standardised audio quality metrics are required to objectively measure the technical attributes of speech to determine when ASR processing should occur and assist with assessing the degree of confidence that should therefore be placed in the output. If technical quality could be standardised and measured it could eventually be incorporated into LR calculations although this would be extremely complicated to realise at a practical level. It is argued that minimum standards of speech quality are required for capture and ideally these would be in line with acceptability criteria and agreed across organisations. Recently it has been extremely encouraging to see the Audio Engineering Society (AES) running regular articles and conventions drawing attention to forensic audio analysis, particularly in the context of audio capture, enhancement and speaker comparison/ID.

Finally, process(es) for audio submission should be managed to ensure that the highest quality recording (i.e. at point of capture) is assessed rather than a transcoded/degraded copy and that the operator is empowered to determine the suitability for ASR comparison with acceptance that a high proportion of speech files submitted for analysis are likely to be rejected. If ASR systems are to be used more extensively in the future, Morrison et al. (2016) and Gold and French (2011; 2019), then greater consensus pertaining to technical standards and acceptability criteria will produce more accurate ASR outcomes and provide greater confidence in speech technology.



## 12.2 Practical Recommendations

As discussed over the course of this thesis, when transitioning theories derived from research conducted under controlled conditions to in-field application it is acknowledged that audio/speech files submitted for speaker comparison analysis will most likely contain similar but often much more complex combinations of acoustic contaminants and inhibitors. File origin (provenance) and the end to end audio signal path are frequently unknown. Nevertheless, in considering the conclusions drawn from each of the chapters in conjunction with research completed throughout the duration of this thesis practical recommendations are offered for consideration.

### 12.2.1 Net Duration Recommendations

The experiments demonstrated that using speaker models of significantly varying length can cause an ASR system to become unstable, with shorter net duration speaker models tending to produce more FP and FR results. A practical recommendation to mitigate for this would be the management of speaker models to provide more uniform duration across the dataset, i.e. capping duration and non-enrolment of low duration speech samples.

Performance tipping points were reached when the total net duration of speech compared fell to less than a minute with a cliff edge noted at <10s for either speaker model or the test audio on the i-vector system (7.13). It is recommended that minimum operating thresholds should be conservatively set and thresholds applied to prevent comparisons occurring on very short duration samples at all.

As stated, net speech duration and audio quality are highly interconnected. Varying proximity (to the microphone), faults, signal drop-outs and/or environmental noise (e.g. music, overlapping speech, babble etc.) is likely to cause lower net duration to pass the speech detection phase and experiments showed that switching speech detection off or lowering the SD threshold to force acceptance could assist – however, it is recommended that much care would be required in doing so to avoid non-speech contamination.

For higher net duration (approximately 1m SA to 1m TA) experiments showed that a likely plateau for performance was forming, suggesting that adequate speech data was captured in the statistical model(s) (0.005% EER on the i-vector system). Nevertheless, as maximum duration was not specifically tested in the experiments, detailed recommendations in that regard are not specifically made. In addition, a counter argument might suggest that the inherent variability of speech (e.g. different languages spoken, mood etc.) are likely to benefit with additions to a speaker model (beyond 1m). Alternatively, another suggestion might be to create different speaker models for the same speaker, each totalling a fixed length (e.g. different acoustic conditions/channels and languages etc.) even using different/switchable normative datasets. The area of combining data from different

sessions/conditions/languages etc. for the same speaker, to identify portions of the statistical model pertaining to those criteria/channels, is currently progressing – see sections 7.3, 13.7 and WCCN.

In summary, there are many variables that can significantly reduce the quantity of viable net speech either being captured or passing the detection phase (VAD). Practitioners should have a strong understanding of where ASR performance can be influenced by net duration and the extent to which performance can be degraded. It is hoped that the new data provided will assist in providing guidance and it is recommended that new tables are created by the practitioner on their own ASR systems and redrawn as they are updated/evolve.

Finally, in 2013, the net duration of speech samples was placed under scrutiny in the George Zimmerman trial (U.S.). Figure 12.1 provides a summarised section of the transcript pertaining to net duration and the use of an ASR system. The methodology was questioned in court by the defence as advised by expert witnesses G. Doddington, J.P. French and H. Nakasone. It was determined that practitioners should not duplicate sections of speech to meet speech sample threshold(s) for ASR use. It was also stated that net duration was only one factor that was heavily disputed and the enrolment of non-modal voice (in this case screams) was also strongly criticised by the defence.

**Figure 12.1:** Zimmerman trial (2013) transcript summary (next page)

[From: legalinsurrection.com/2013/06/zimmerman-prosecutions-voice-expert-admits-this-is-not-really-good-evidence/](http://legalinsurrection.com/2013/06/zimmerman-prosecutions-voice-expert-admits-this-is-not-really-good-evidence/)

Names and ASR manufacturers have been redacted. Double line = transcription break.

## **Practitioner Admits Audio Recording Too Short for Analysis**

**So He Loops It.** One of the most shocking moments of the testimony, however, involved (the practitioners) odd methodological choices, in particular his response when his (speaker verification software) informed him that the screams were of too short a duration to enable analysis.

**Defence Lawyer:** “What was the total duration of those 10 screams?”

**Practitioner:** “About 7 seconds.”

**Defence Lawyer:** “I believe you earlier testified that you want 16 seconds of speech for the software to work reliably.”

**Practitioner:** “The software would like to see 16 seconds. If you put in 7 seconds the software says it’s not long enough. So I doubled it up. I repeated the same audio twice, back to back, and then the program would run the analysis.”

**Defence Lawyer:** “Are you aware of studies or research of this software using that method you describe where you enter less than enough speech by doubling up.”

**Practitioner:** “You don’t enter less speech by doubling up, you increase the amount of speech.”

**Defence Lawyer:** “So, if I say ‘Hello’ and repeat that 16 times, is that 16 seconds of speech that’s appropriate for your software?”

**Practitioner:** “No, [stuttering] as far as doubling up it just provides enough to make a decision.”

**Defence Lawyer:** “So in other words unless you had doubled it up and looped them the software would have rejected the sample.”

**Practitioner:** “A screen comes up and tells you that it won’t run. We didn’t have any more words to give the machine, so I doubled it up, because that’s all we had.”

**Defence Lawyer:** “Because it then thought there was more speech than it previously had.”

**Practitioner:** “It knew there was more speech.”

**Defence Lawyer:** “When you realized that there was a problem with the speech sample, besides all the other problems, it wasn’t long enough for the machine to be able to conduct a reliable analysis, is that correct?”

**Defence Lawyer:** “It wasn’t long enough for the machine to do its analysis.”

**Defence Lawyer:** “So at that point you contacted someone from the company to ask them what to do?”

**Practitioner:** [Angry] “I’ve been using this software for 15 years.”

**Defence Lawyer:** “I thought you said only 1 to 2 years?” “You’ve never done this before, with this software, by looping the unknown sample until you have enough duration to put into the machine?” “So this is brand new stuff, isn’t it?”

**Practitioner:** “Brand new about the looping.”

**Defence Lawyer:** “So, you needed 16 seconds, and you found 7.”

**Practitioner:** “The software recommends 16 seconds.”

**Defence Lawyer:** “You knew that was half of the recommended minimum speech sample.”

**Defence Lawyer:** “So do you think that the 20 word minimum means the same word repeated over and over, or are you looking for something phonetically balanced?”

**Practitioner:** “What’s your definition of ‘phonetically balanced’?”

### 12.2.2 Signal to Noise Ratio Recommendations

The experiments demonstrated that lowering SNR adversely influences ASR performance. For the i-vector system tested a tipping point was identified at the addition of noise at -20db/-15db where the EER% exponentially rose. At the addition of 0db noise, the system effectively failed with EER 45.5% (white noise) and 48.3% (pink noise) on matched conditions. One explanation for this is that at the feature extraction stage the system can effectively no longer discriminate between speech and noise at all so the statistical model is polluted producing high speaker confusability.

Results showed that tonal (pink noise) was found to influence ASR performance to a marginally greater extent and this is likely due to the lower frequency content (higher noise on F1, F2, F3). Whilst not tested, it is suggested that it is likely that babble (e.g. background speech from non-speakers) would degrade ASR performance further as found in Desphande and Holambe (2011a) - since babble is more similar to (foreground) speech than white or pink noise. It would be strongly recommended that the type(s) of noise present should therefore be considered by the practitioner, perhaps in conjunction with (WADA) SNR estimation.

Recent research by Al-Noori and Duncan (2019) examined the incorporation of training data with different noise profiles (babble) added to minimise the channel mismatch between SM and TA. Their research was conducted using a bespoke corpus of 110 speakers (55 male, 55 female) recorded in anechoic conditions and the addition of TIMIT data (3 different types of babble) using a GMM-UBM ASR system. Results showed a similar performance tipping point to the experiments completed in this thesis when the SNR is reduced below 15db. Although a marginal improvement in EER% was obtained when degraded training data was incorporated, this was not the case for >20db SNR where they pointed out that ASRs tend to perform better without data augmentation. A repeat of the experiments is suggested using a state of the art i-vector system to establish if the addition of degraded training data (e.g. speech 'babble') can improve ASR performance or whether the performance improvements observed on the GMM-UBM system tested could be achieved by upgrading to an i-vector system.

Godin, Sadjadi and Hansen (2013) suggested certain noise reduction techniques could improve ASR performance – however, they found that noise reduction also degraded ASR performance and, in some cases by a considerable amount (2013: pp.3658-3659). Their study also found that different techniques would be required for ideal/matched and non-ideal/unmatched conditions (2013: p.3659) as well as for GMM-UBM and i-vector ASR systems (and now likely x-vector/DNN systems).

Audio enhancement, prior to ASR analysis, was also studied by Künzel and Alexander (2014). They added different types of noise (e.g. music, babble, road traffic) to the speech of 10 speakers at different SNR settings and then analysed output using an Agnitio Batvox system both pre and post

audio enhancement. The audio enhancement system used was CEDAR ([Cedaraudio.com](http://Cedaraudio.com)). Künzel and Alexander concluded that audio enhancement, in some instances, could improve ASR performance (particularly in the case of music and moving vehicle). However, they also reported that in three instances audio enhancement degraded ASR performance. In other cases they assessed that enhancement wasn't necessary, based on the very negligible ASR performance decrease noted when SNR >6db (2014: p.251).

In summary – although all studies show promise, investment in research is required if audio enhancement techniques are not to inadvertently degrade ASR performance. It would therefore not be recommended that attempts to remedy poor SNR with the use of audio plug ins pre ASR analysis was completed without a very high degree of technical expertise in all the relevant fields. As previously discussed (12.1.1) other recommendations relating to SNR include:

- i. Avoiding the comparison of very high SNR and very low SNR recordings;
- ii. WADA SNR estimates could be obtained on enrolment to identify potential SNR issues;
- iii. SM or TA files should not be populated with very high SNR and/or very low SNR (consider separate speaker models for the same speaker).

### **12.2.3 Reverberation Recommendations**

Experiments showed that even relatively small reverberant spaces (RT60 of .30s) negatively influenced ASR performance with EER% and Cllr (accuracy) degraded. Larger reverberant spaces were more detrimental to performance with a tipping point of approximately T60 >1s, dependent on the complexity of reflections. The most plausible explanation for this is that the smearing of speech data in the time domain affects the feature extraction stage (which operates in subsecond frames) and this, in turn, decreases the detail captured in the statistical model(s). It is therefore recommended that speech files are assessed by a technical specialist for reverberant noise and that reverberant and non-reverberant files are compared with much caution. Combining speech data recorded with different reverberant conditions into the same speaker model(s) would not be recommended without fully understanding the influence on speaker/ASR performance. It would also be recommended that any reporting resulting from comparing speech files which contain reverberant speech are adjusted accordingly with respect to confidence.

Audio enhancement techniques, to remove reverberation, could potentially assist (Wu et al., 2017; Guzewich and Zahorian (2017) but more research is clearly required before recommending it as part of the ASR workflow. For example, it cannot be assumed that applying a similar, simulated reverberation to 'force a channel match' would be successful. Reverberation is extremely complex and further research is also required to establish how well 'real world' reverberation can be successfully modelled.

## 12.2.4 Frequency Bandwidth Recommendations

There are clearly many factors which should be considered when completing technical assessments of speech files with constrained frequency bandwidth submitted for comparison. Experiments demonstrated that examination and suitability of frequency bandwidth/sample rate of normative data should also be completed to avoid mismatch between the SM/TA and normative data. For 1 to N comparisons it is also recommended that the technical quality of all speaker models should ideally be uniform in terms of frequency bandwidth to prevent the skewing of results. It is conceded that many ASR systems constrain the frequency bandwidth on enrolment, to assist with achieving uniformity. Nevertheless, the system cannot compensate for the enrolment of files which are below the sample rate threshold set on the ASR (i.e. 0-3.5kHz, 07kHz SR).

## 12.2.5 Transcoding/Codec Recommendations

Performance tipping points were noted when lossy codecs at high data compression settings were used for transcoding (e.g. MP3 at <32kbps). One explanation for this is that the loss of data prior to the feature extraction stage degrades the efficacy of the statistical modelling phase. From analysis the performance degradation appears more noticeable in the upper frequency range with formants F3 and F4 effectively degraded and high F2 affected (for heavy data compression) for which data gaps were also visible in the spectrogram (Figure 11.29). Artificially elevated scores were noted for certain compressed files and results from experiments supported Silovsky, Cerva and Zdansky (2011) that, in some cases, bulk transcoding (SM and TA) caused an ASR performance improvement.

Transcoding can clearly cause variable ASR performance – and it is therefore recommended that speech files are carefully inspected for transcoding damage, particularly where the full provenance of a file is unknown. If the codec history of a submitted file is known it is recommended that it is fully documented and factored into analysis. If proceeding to comparison, using an ASR system, reporting should be notated/caveated accordingly with the confidence of assessment adjusted for transcoded files. It is recommended that kbps, in and of itself is not applied as an acceptability criteria (see 12.1.3, point iv). It is also recommended that speaker models that have passed through different codecs are carefully assessed before integration on the same ASR system within the same comparison set(s), since those speakers which are effectively damaged from data compression have been shown to be more prone to FP/FR outcomes.

Finally, if an ASR is used on a computer network it is recommended that files are not transcoded as they are transferred or, if they have to be, a lossless codec is applied which is assessed by the ASR/speech analyst as appropriate. It is hoped that the new data that has been collected from the transcoding experiments in this thesis assists with informing that process.

## **12.2.6 Control Corpora and Test Data**

Investment has been made by IAFPA, NIST/SRE, LDC and – more recently - Morrison and Enzinger (2019) to provide access to control corpora to test ASR systems and assist with optimisation and establishing baseline performance. Whilst there are cost and data sharing implications, the widespread availability of corpora will allow greater consistency of measuring ASR performance, improve understanding of speaker performance with reference to the demographics of the corpora and establish baseline positions for different ASR systems. It is therefore recommended that wider access to more control corpora, including degraded data, is granted to assist with calibrating systems, validating methodology and ensuring systems are consistent in output.

## **12.2.7 Automatic Speaker Recognition System**

### **Recommendations**

ASR systems are continually evolving and this became very apparent during the course of writing this thesis. Completing research experiments over a long time frame, relative to ASR development, (approximately 6 years) saw 12 different iterations of the ASR software including core architectural design changes (GMM/UBM to i-vector). There were also 4 different versions of the iZotope editing software and 5 iterations of Bio-Metrics. It was interesting to note the overall progress in ASR performance, with initial preliminary tests on baseline data producing EER% scores of around 6% reducing to 1% with the later i-vector technology experiments producing .0051% EER.

Constant updates presented a risk in terms of additional variability. To mitigate for this, when updates were applied, models were re-enrolled and results cross validated to ensure that output was consistent and that patches and changes did not introduce further variability. Overall, the perpetual ASR evolution emphasised the importance of regularly updating systems to maintain pace with evolution, however this cannot be recommended enough in terms of ASR performance. Updating systems has resource implications, not only in terms of purchasing upgrades, but relating to the retraining of speaker models, retention of data (since re-enrolment is essential) and training costs. It is therefore recommended that careful system version and data control becomes an integral part of the workflow. ASR system changes should be carefully documented and changes fully tested using specific control corpora to measure EER% and Cllr.

Finally, the constant evolution of ASR systems and infinite combinations of settings and normative data is likely to generate more challenges, rather than less, for proceeding ASR output to court - i.e. repeatability, reproducibility and consistency of EER%. It is therefore recommended that, where possible, standards accommodate for rapid change and that they are frequently updated.

### **12.2.8 Performance Metrics**

From the experiments completed, recommendations are made pertaining to ASR system performance metrics in that they should not be solely represented by an EER% figure and that accuracy (Cllr) should be incorporated into reporting. It is recommended that both EER% and Cllr are measured for the specific ASR system and conditions. Experiments have also demonstrated that LLR score separation and score height (H0 and H1) provide useful performance information and can mitigate against high FP, low TP instances by identifying those speakers/audio files which can be problematic in an ASR system. Finally, zoo plots, are recommended to ensure ASR systems are performing correctly overall and to assist with identifying outlier speakers and/or other issues regarding acoustic variability.

### **12.2.9 Auditory Analysis**

In the context of acoustic variability ASRs have been shown to be an important and effective analysis tool but not definitive in output. It is therefore recommended that auditory analysis, by a trained practitioner, should continue to independently validate ASR output – particularly in cases of high importance and potential transition to evidence.

## **12.3 Should Automatic Speaker Recognition Transition to Forensic Use?**

In the context of acoustic variability, should ASRs be used in forensic speaker comparison (FSC) work and in what capacity? How should we address the complex and enduring issues such as quantifying variability and selecting appropriate population data and ASR settings?

The thesis has highlighted many technical difficulties in transitioning ASRs to FSC but also many positive aspects relating to accuracy. In many instances, the i-vector system was shown to be remarkably resilient with regards to acoustic variability.

An additional concern could be that an ASR system is somehow viewed as a replacement for the expert. It certainly does not currently fulfil that role. Across the international community, opinion is somewhat divided. One view is that ASRs are not ready to transition to evidential use yet and that they should firmly reside in the investigative domain (where they are useful for exploring large data) and auditory phonetic analysis must underpin results. A counter opinion would be that provisioning common systems, providing improvements to normative data/audio capture and developing ASR expertise could meet the necessary standards regarding repeatability and reproducibility. Yet all options require a significant investment in resources and central governance, currently lacking in the U.K.



To return to an initial aim of this thesis regarding examining the suitability of ASR use in FSC it has been demonstrated that acoustic variability can have a significant bearing on ASR reliability such that this effect alone could render the output unreliable for use in evidence. However, if used in a highly controlled manner on audio that has been assessed by a technical expert as suitable then output could provide an accurate likelihood output. ASRs are, currently, ideally suited to investigative (batch analysis) and pre-forensic casework where they can provide an alternative and empirical view on the data presented for comparison. Although, to most experienced forensic speaker comparison experts, traditional methodology would be more preferable - a workflow which combines ASR analysis with acoustic and phonetic analysis would be more comprehensive.

Whilst there is no current UK precedence for the presentation of ASR system output either in conjunction with auditory analysis evidence or in isolation it is suggested that systems and process/workflow are not yet regulated enough to make this progression from investigative use to evidential use – although this is changing (ENFSI and the UK Forensic Science Regulator).

In a recent study examining the wider issues surrounding the admissibility of forensic voice comparison testimony (in Australia), Morrison (2018: p.23) states support for ‘empirical validation irrespective of approach’. In a later study Morrison and Enzinger (2019) state that whilst recommending that their ASR performance results are not applicable to other conditions, cases and systems, they encourage practitioners to support their forensic voice casework (i.e. auditory analysis) with empirical ASR output (2019: p.38). These are eminently sensible recommendations and resonate with the experiments completed in this thesis. It is certainly more practical to measure ASR system performance than a human auditory practitioner. Unfortunately, as demonstrated, acoustic variability, normative data and ASR settings can produce a wide range of empirical ASR output. The operator themselves are also effectively a part of the system – making decisions, selecting data and choosing system settings. So, empirical validation is certainly to be aspired to but difficult to achieve.

In their recent analysis of ASR systems on real forensic data, Solewicz et al. (2012: p.86) summed up simply that systems can be a valuable support in decision making for the forensic examiner. Their findings also supported the view of the wider community that automatic speaker comparison technology is generally less than perfect, can require significant human assistance to be utilised in any meaningful forensic context and that ASRs should be used – but with caution. In addition, Solewicz et al. (2012) emphasised the difference when applying laboratory standards to real forensic case data – which can be unrealistic due to the acoustic variability of the latter.

‘The typical attributes of forensic material sometimes lead to unpredictable results that are not necessarily consistent among the systems investigated.’ Solewicz et al. (2012: p.90).

Throughout the writing of this thesis the infinite variability of speech, the inconsistency of acoustic conditions, the wide range of technical quality of recordings, the complexity of assembling population data and the wide variety of ASR systems (and settings) demonstrate how complex it will be to transition ASRs to forensic application.

To summarise, it is intended that the research experiments completed provide useful data to assist in informing users about acoustic variability and assist with understanding the associated risks. Finally, key issues pertaining to ASR use will endure in respect of reliability and reproducibility in the context of:

- i. Agreement on normative data and suitability, particularly for acoustic variability and mismatched conditions;
- ii. Validation of results across multiple laboratories – e.g. different systems/normative sets ASR make/model, underlying architecture, configuration/settings and thresholds;
- iii. Measurement pertaining to audio quality/quantity;
- iv. Agreement on acceptability criteria – which likely requires adapting to (i) and (ii);
- v. Agreement on metrics regarding ASR performance;
- vi. Explanation of LR/LLR and ASR output to non-specialists/the courts.

There are, of course, other issues outside the scope of this thesis - such as the current inadmissibility of intercepted audio (i.e. mobile and telephony) in the UK courts, in comparison to eavesdropping audio.

This chapter discusses opportunities for further research in reference to the experiments completed.

### **13.1 Combining Acoustic Conditions**

It is recommended that research should examine the combining of acoustic variability (e.g. reverberation, SNR and transcoding) with the objective of determining if there are any broad mathematical relationships to be drawn in respect to degradation and EER% and Cllr. It is expected that inhibitors and contaminants will have a cumulative effect on ASR performance when combined, but the sum of that effect cannot easily be predicted. It is hypothesised that performance degradation is not likely to be linear and will probably introduce further cliff edge effects. This research would ideally require much larger scale data modelling to better represent hundreds, if not thousands, of different permutations of acoustic variability. Research could assist with assembling normative datasets and compensating for non-ideal/mismatch of acoustic conditions.

### **13.2 Modeling Automatic Speaker Recognition System Environments**

Modelling using extremely large datasets of SM and TA and simulating permutations of ASR systems (and settings) under thousands of different conditions could provide more reliable estimates as to ASR performance. Parsing much larger datasets of artificially degraded/modelled normative data could potentially produce more reliable estimates as to the degree of error in output arising from acoustic variability. The data could also better inform ASR performance under complex conditions, assist with population data selection or guide as to the inherent variability in the ASR itself. It could also produce optimum settings for analysis of very complex acoustic conditions.

#### **13.2.1 Applying Big Data for Mismatch Compensation**

The lack of calibration data for addressing channel mismatch is an enduring issue - also highlighted by Morrison, (2018b). Aside from completing a controlled degradation of either SM or questioned audio (or both), one solution for better compensation with respect to mismatched conditions could be in the creation of recording adapted background models (RABMs). This was originally proposed by Becker et al. (2010) in relation to improving performance for ASR systems on real case data (747 recordings, from 184 speakers). The team showed EER% benefit (from 17% to 7%) applying this method. They suggest that future success would be dependent on the collection of very large volumes

of relevant population data and adaptation of the background dependent on the comparisons completed.

Degrading large batches of normative data was more recently proposed by Ferràs et al. (2016) who demonstrated that adding a database of over 60 hours of environmental noise and 100 impulse responses to simulate conditions could improve the matching of conditions between SM and questioned audio. Their experiment produced an improvement on a bespoke i-vector system of between 40% and 80% relative EER (Ferràs et al., 2016: p.530).

In summary, all these methods show promise and the opportunity to develop extremely large data sets featuring millions or hundreds of millions of audio files, specifically to compensate for mismatch, should be researched. Nevertheless, simulating accurate and complex channel conditions will not be without difficulty. In addition, there are also ethical implications - see U.K. General Data Protection Regulations or GDPR (2018).

### **13.3 Pre-Analysis Audio Enhancement**

As previously discussed (12.2.2) the sparing use of audio enhancement was shown to marginally improve ASR performance in some instances – although it also degraded ASR performance - consistent with Künzel and Alexander, P. (2014). Audio enhancement prior to ASR analysis is quite controversial in terms of digitally altering the audio which, it could be argued, would be applying software to potentially change/elevate ASR LR output. From the small scale experiments completed, it is hypothesised that most noise reduction techniques are unlikely to significantly improve ASR performance as in most cases the degradation is due to the loss of speech data, sometimes referred to as ‘moth holes’, which can not simply be filled (spectrogram, Figure 11.21). In addition audio enhancement will introduce an additional and undesirable set of unknown (new) acoustic variability.

The many approaches to audio enhancement were not within the scope of this thesis – but removal of predictive and variable noise such as music (reference cancelling), adaptive filtering (removal of noise which adjusts to the incoming signal) and the filtering of specific frequencies (low pass, high pass, band pass and comb filters) are generally considered the most popular techniques. For now, however, audio enhancement pre-ASR analysis should firmly reside in the investigative only application of ASRs and further research is recommended.

### **13.4 Feature Extraction Methods and System Fusion**

Recent research from Athulya and Sathidevi (2018) demonstrated that ASR system performance degradation caused by codec distortion can be partly compensated for by applying an alternate feature extraction method and fusing the output with output from another feature extraction method.

In exploring this, the authors used Power Normalized Cepstral Coefficients or PNCC and then applied fusion with traditional MFCC feature extraction on a GMM-UBM system. The TIMIT corpus consists of 630 speaker and 80 speakers were used for testing (SM and TA) with the remainder used for UBM/normative purposes. They noted a reduction in EER from 22.4% to 2.5% (optimum baseline 0.3165% EER) – showing much promise in this area. Another recent study by Fedila, Bengherabi and Amrouche (2018) also found performance benefit from a similar approach - fusing a Gammatone Product-Spectrum Cepstral Coefficients GPSCC and a MFCC GMM-UBM system tested with TIMIT speech files degraded using the G. 722 codec (Figure 13.1).

**Table 13.1:** Fedila, Bengherabi and Amrouche (2018: p.16734, Table 6)

EER% performance using GPSCC and MGCC features and their score level fusion at different bit-rates					
Features	6.60 kbps	8.85 kbps	12.65 kbps	15.85 kbps	23.85 kbps
GPSCC	31.70	24.29	13.61	<b>9.05</b>	4.69
MGCC	37.15	22.00	10.00	10.10	<b>3.85</b>
Fusion	<b>31.01</b>	<b>21.68</b>	9.68	9.60	4.15

In the preliminary tests which experimented with LTFD measurement as an alternative feature extraction method it was noted that, whilst overall EER% was generally not as good as either MFCC GMM-UBM (or MFCC/i-vector), LTFD did appear more resilient to noise than the MFCC GMM-UBM system. It is suggested that this could be a specific area for further research - i.e. could there be benefit from fusing a new, improved LTFD extraction method with the x-vector/DNN approach?

Adding other speech measurements, including those more determined by content (i.e. text dependent/output from speech to text to extract similar entities) could also be potentially fused together for further performance benefit. In the context of the experiments completed on frequency bandwidth - combining inverse MFCC with the filters spaced in reverse and additional detail in the upper frequency band, this could also provide additional performance benefit and requires research.

Schieland and Zitzelsberger (2018) also evaluated different methods of formant tracking and proposed that Deep Learning (DL) offered the best solution, conceding that LPC trackers were approximately twice as imprecise as humans, that the trackers must be adjusted for gender (2018: p.2847) and the large requirement for marked data (for DL). It is recommended that these studies form the basis for further research to use DL formant tracking, in the context of ASR performance and acoustic variability.

Finally, the fusion of score matrices, i.e. multiple tables of results from different systems, is appearing within software such as OWR Bio-Metrics 2018 (Alexander et al., 2018). This could effectively bring different ASR system output together more easily – utilising the best feature extraction and statistical modelling methodologies of multiple systems to provide additional EER% and Cllr benefit, particularly if common standards of ASR output are observed.

### **13.4.1 Automatic Speech Recognition**

Recent research by Fujitsu (2017) on text dependent low duration utterances applied machine learning algorithms for both speaker verification and speech recognition. It is not noted which corpus was used or the diversity of the speaker population (language, gender, age etc.). A typical text dependent use case could apply to authorising voice passwords (e.g. account access for telephone banking) where phrase repetition could assist in offsetting low duration. The Fujitsu approach showed an EER of 2.2% on utterances (<3s) on a set of 200 speakers (no further details supplied). This broadly supports the hypothesis that ‘what is said’ is of high significance to verification performance. It is then likely that the improvements in ASR performance will occur i.e. when the trained utterance and questioned audio are identical. Whilst this could work well in the context of banking etc. performance figures achieved in the text dependent domain obviously cannot be applied to forensic application. It is also suggested that false positive rates would rise and EER% is not likely to remain low if increasing the candidate pool (from 200). Whilst controversial, a machine learning approach like this could potentially assist with much wider speech to text and speaker recognition problems (e.g. large scale) – for example, applying machine learning re identical/similar phonemes, words or phrases which could then be extracted from big data (e.g. to create multiple speaker models) – this might then improve similarity of utterance(s) with the questioned audio.

### **13.5 Automated Audio Quality Measurement**

Assessing acoustic variability and audio quality objectively, quickly, at scale and in a repeatable and reproducible manner is likely to become more important in the context of acceptability criteria and common standards. New techniques for extracting audio quality metrics from audio files using semi-automatic processes requires further research. For example, the output of audio quality measurements (.xml files) could be incorporated into the diarisation process to better determine which sections (or audio files) are suitable for ASRs (or not).

### **13.6 Alternate Approaches to Speaker Model Generation**

An area of interest that arose during research was the concept of using WCCN to better separate/measure channel influence through the use of multiple speaker models recorded from

multiple sessions. As previously discussed, the system is trained to know that multiple reference samples are from the same speaker. Then i-vectors pertaining to speaker specific information can then be better separated from those that relate to channel information. Further research is required in this area.

## **13.7 X-Vector Automatic Speaker Recognition Systems**

As discussed, the x-vector DNN approach (Snyder et al., 2018) was developed towards the end of writing this thesis and ASR manufacturers have been quick to exploit the improvements in statistical modelling density and apply deep neural network (DNN) techniques to comparison. Several new x-vector ASRs have recently been launched and a beta version of xVocalise was recently used in the context of the multi-laboratory, forensic evaluation trials (Morrisson and Enzinger, 2019; Alexander and Kelly et al., 2019). Ten different ASR systems were tested in total using a special forensic evaluation corpus referred to as ‘forensic\_eval\_01’ (Morrisson and Enzinger, 2019: p.37).

In summary, it was noted that the 3 x-vector PLDA/DNN systems that were tested marginally outperformed their i-vector/PLDA predecessors (e.g. iVocalise 0.07 EER%, Cllr mean 0.23 and x-Vocalise 0.05 EER%, Cllr mean 0.213). Further research in this area will be important in terms of quantifying improvements between i-vector and x-vector architecture, particularly in relation to acoustic variability.

# Chapter 14 Conclusion

---

This thesis demonstrates the importance of acoustic variability on ASR performance through the investigation of five acoustic conditions. The experiments completed contribute to the field through estimating the extent of ASR performance degradation, highlighting the importance of acoustic variability, emphasising the significance of completing full technical assessments (of recordings), raising awareness as to the risks surrounding acoustic variability and providing recommendations to mitigate. In addition, further opportunities for research pertaining to acoustic variability and ASR systems were identified and presented.

Specific contributions to the field were made through the provision of relevant new data. It was demonstrated that:

- i. Performance tipping points were identified, these should be measured and known for a specific ASR.
- ii. Performance benefit can be gained by increasing the sample rate for ASR analysis to 22kHz SR, 0-11kHz frequency bandwidth (i.e. WB). This assumes that other acoustic degradation is not present and the ASR system would require optimisation accordingly (feature extraction process and normative data).
- iii. ASR systems can be successfully applied to assessing speech files less than 8kHz SR, 0-4kHz frequency bandwidth (i.e. less than NB) – but accuracy and performance is degraded and much care must be taken to optimise the system, calibrate/match conditions. Reliability of the system and output should be measured and incorporated into reporting.
- iv. Constraining the frequency bandwidth beyond 0-3.5kHz, 07kHz SR (on male speech) was shown to be (likely) reaching the theoretical limit of extractable speech data for which accurate ASR is possible. The experiments broadly supported the minimum high frequency standards as surveyed by Gold and French (2019) - see 2.1. Female, child speech and tonal languages were not tested but the experiments show that constraining F3, F4 and above influences ASR performance and so it is likely that the frequency range would need to be adjusted accordingly for female and child speakers.
- v. It was consistently demonstrated that the i-vector ASR outperformed the GMM-UBM ASR. This is in line with other research and supports the updating of ASR systems.
- vi. The minimum acceptance criteria for net speech duration for some practitioners was noted as 3 seconds (Gold and French, 2019) (2.1). From the results in chapter 7 it would be recommended this is too low and minimum acceptance criteria should be revised upwards (assuming ASR use, rather than auditory analysis).
- vii. Combining mismatched acoustic conditions within the same set of comparisons was shown to provide poor overall output (e.g. mixture of high/low true positives and false negatives).



Matched conditions, between the question and reference samples, were consistently shown to improve ASR performance with the exception of net duration where minimum quantities were identified (Tables 7.9 and 7.10).

- viii. Recommendations are made in relation to the management of speaker models including, where practicable, more uniform conditions to prevent the skewing of results (7.7).
- ix. Controlled degradation (forcing matched conditions), where carefully applied by a skilled practitioner, could restore ASR performance but caution is recommended and further research is required.
- x. Audio enhancement was shown to have both a positive and negative effect on ASR performance and should not be used without much caution. More research is required.
- xi. Larger datasets for the population/normative data and the addition of reverberant material provided an ASR performance improvement when comparing material containing speech with reverberation.
- xii. Removal of VAD, in some instances, assisted with ASR accuracy for degraded speech files but should be used with much caution.

As ASR systems become more widespread in their application one concern may be that the quality and standards surrounding their usage is diluted. This could occur through the lack of expertise to technically assess audio, operate the ASR correctly and governance. With large scale data enrichment services becoming available on cloud services, to commercial organisations (e.g. call centres/account verification) it is also highly likely that speech technology enrichment services will occur with even less visibility to either the end user or a system administrator.

Despite the enduring development of ASRs, issues were highlighted around the implications of conducting ASR analysis on degraded audio and the relative immaturity and current suitability of ASRs to directly transition to forensic application. Broader recommendations have been raised suggesting strategic requirements for investment, consensus within and across organisations as to systems and processes in addition to methodology and the importance of collaboration regarding expertise and data. In discussing use cases it is important to differentiate the spectrum of applications from ‘call-centre ASR systems’ to investigative and forensic capabilities. The latter of which clearly demands much higher ASR performance, standards, governance and issues surrounding channel mismatch (e.g. interview and telephony) – and yet the Netherlands and Germany are succeeding where the U.K. is not. Nonetheless, it could be argued that audio generated to one set of standards might later be required for investigative/forensic use.

As a result of the experiments conducted in this thesis it is hoped that attention and investment is also drawn to the importance of high quality audio capture and consideration of the transmission, reception, networking, archiving and pre-processing of audio prior to ASR analysis. Objective

measurements of audio/speech quality will be difficult to find consensus on and perpetual updates to ASR software and normative datasets, whilst iteratively improving EER%, can exacerbate difficulties relating to transitioning ASR systems to forensic use. The thesis also discussed the sizeable investment and consensus between multiple organisations that will be required to transition ASRs into any part of the forensic process. Efforts to improve our understanding of intrinsic and extrinsic variability must be continued. Large scale investment in marked data for machine learning, better population data (with metadata) and new research will be vital.

As practitioners we should complete stringent qualitative technical assessments, apply sound forensic methodology, validate our results on more than one ASR system and understand/quantify the expected performance bandwidth of our ASR system(s). Not forgetting, that the physiological dimensions of the vocal tract are not as varied across human beings in comparison to DNA, for example. These are all important steps to ensuring that speaker verification using ASRs does not descend into an unreliable pseudo-science as the technology becomes more prevalent.

In summary, there are really three simple options available pertaining to acoustic variability - to either adapt the ASR system and/or modify the audio files or to not proceed with ASR comparison on the grounds of insufficient data.

If applied in a controlled way and with highly validated and robust methodologies, by fully trained experts, ASRs can provide an invaluable investigative tool and should certainly be applied to pre-forensic casework. With careful application, fully trained operators, safeguards and governance in place, the final recommendation of this thesis is that ASRs should eventually, one day, progress to evidential use in the U.K.

# Appendices

---

The following Appendices (A-J) are presented.

- Appendix A** Additional results from Transcoding experiments
- Appendix B** Examples of Preliminary Tests. Examples of VQ analysis method
- Appendix C** Preliminary Tests. VQ and zoo plot results and analysis
- Appendix D** Preliminary Tests. Zoo plots (DyViS, Pakistani and Yorkshire accented data)
- Appendix E** Zoo plot and LR plot .gif animations (individual frames/.bmp)
- Appendix F** Reverberation conditions, Waves IR-L
- Appendix G** OWR Vocalise and iVocalise ASR system specifications and versions
- Appendix H** Slides presented by the author at 2014 IAFPA (Zurich)
- Appendix I** Poddar, Sahidullah and Saha Tables. Net duration research to date (2015)
- Appendix J** Test tone experimentation tables
- Appendix K** Image from Speaker 012, spectrogram analysis
- Appendix L** Response from Parallels re VM software

# Appendix A

---

Additional results from CODEC Testing (Chapter 11).

## **G711 uLaw 8kHz from 22kHz**

0.0337% EER: FARR/FRR: 1, 0: 0.1, 0: 0.01, 2

Mean of H0: 48.8036: Mean of H1: -67.15491

Standard Deviation of H0: 12.94604 Standard Deviation of H1: 27.70791

## **GSM 8kHz from 22kHz**

0.3418% EER: 1, 0: 0.1, 1.33: 0.01, 2.33

Mean of H0: 50.17181: Mean of H1: -54.77997

Standard Deviation of H0: 12.80004 Standard Deviation of H1: 26.22543

## **Speex 32 CBR 16kHz from 22kHz**

0.0286% EER: FARR/FRR: 1, 0: 0.1, 0: 0.01, 2.33

Mean of H0: 47.49995: Mean of H1: -66.24854

Standard Deviation of H0: 12.50942 Standard Deviation of H1: 27.49063

## **G711 uLaw 6kHz from 22kHz**

1.0859% EER: FARR/FRR: 1, 1.67: 0.1, 5.33: 0.01, 26.35

Mean of H0: 58.01739 Mean of H1: 6.33455

Standard Deviation of H0: 7.686879 Standard Deviation of H1: 15.69978

## **G711 aLaw 6kHz from 22kHz**

0.6700% EER: FARR/FRR: 1, 0.67: 0.1, 4.85: 0.01, 20.33

Mean of H0: 58.14906 Mean of H1: 6.154152

Standard Deviation of H0: 7.38196 Standard Deviation of H1: 15.52763

## **Dialogic ADPCM 8kHz from 22kHz**

0.3519% EER: FARR/FRR: 1, 0.00: 0.1, 1.33: 0.01, 3.67

Mean of H0: 50.96175 Mean of H1: -56.17345

Standard Deviation of H0: 12.98414 Standard Deviation of H1: 27.22046

## **Opus 16kbps Constrained Variable from 22kHz**

0.0741% EER: FARR/FRR: 1, 0.00: 0.1, 0.33: 0.01, 1.00

Mean of H0: 51.51166 Mean of H1: -58.62491

Standard Deviation of H0: 12.57021 Standard Deviation of H1: 26.44784

**Speex Quality Preset 1 from 22kHz**

1.6515% EER: FARR/FRR: 1, 2.00: 0.1, 13.48: 0.01, 30.78

Mean of H0: 56.0914 Mean of H1: -10.46355

Standard Deviation of H0: 10.20975 Standard Deviation of H1: 20.91397

**Speex Quality Preset 0 [Lowest] from 22kHz**

2.3199% EER: FARR/FRR: 1, 5.00: 0.1, 17.00: 0.01, 32.33

Mean of H0: 73.71233 Mean of H1: 28.96915

Standard Deviation of H0: 7.074686 Standard Deviation of H1: 15.89392

**Ogg Quality Preset 0 Lowest from 22kHz**

0.2845% EER: FARR/FRR: 1, 0: 0.1, 0.33: 0.01, 1.67

Mean of H0: 48.87741 Mean of H1: -65.51038

Standard Deviation of H0: 13.16722 Standard Deviation of H1: 27.09071

**AAC Average Bit Rate 16 from 22kHz**

0.3300% EER: FARR/FRR: 1, 0.00: 0.1, 0.67: 0.01, 3.67

Mean of H0: 65.38939 Mean of H1: -4.177887

Standard Deviation of H0: 8.312696 Standard Deviation of H1: 19.28006

**MP3 Average Bit Rate 8 from 22kHz**

4.3030% EER: FARR/FRR: 1, 14.33: 0.1, 37.48: 0.01 62.01

Mean of H0: 49.51321 Mean of H1: 0.3016596

Standard Deviation of H0: 11.00564 Standard Deviation of H1: 18.12873

# Appendix B

## Examples of Preliminary Tests re VQ data with observations.

Example section of a Table adapted to view VQ data from Stevens and French (2013). The table developed was extremely large, so here just showing data for the first 12 speakers. The total scoring row refers to experimentation seeking correlates against extent of VQ, although this did not take into consideration the rarity of feature.

**Appendix B: Table 1: Example VQ Data, Stevens and French (2013) + subsequent analysis**

	SPEAKER										
	001	002	003	004	006	008	009	010	011	012	
Setting											
Lip rounding/protrusion											
Lip spreading											
Labiodentalization									2		
Extensive range											
Minimised range											
Close jaw											
Open jaw											
Protruded jaw											
Extensive range											
Minimised range											
Advanced tip/blade	1	2	2	2	3	3	2	2	3	2	
Retracted tip/blade											
Sibilance	2	1	2		1				1	1	
Fronted tongue body	2	2	2	2	2	2	2	2	3	2	
Backed tongue body											
Raised tongue body											
Lowered tongue body											
Extensive range											
Minimised range											
Pharyngeal constriction						2	2				
Pharyngeal expansion	2										
Nasal			1		1		3				
Denasal	1	2							2	2	
Raised larynx		2	2			2	2				
Lowered larynx	3			2							
Tense vocal tract											
Lax vocal tract											
Tense larynx			1				2				
Lax larynx	2			1							
Creaky	1	2		3	3	2	3	1	2	4	
Whispery	2		2				1				
Breathy	3		2	3	2	2		2	2	1	
Harsh		1	1		1	1	3	1	1		
Tremor			2								
	001	002	003	004	006	008	009	010	011	012	
Total scoring	19	12	17	13	13	14	20	8	16	12	
MFCC			DOVE			DOVE				DOVE	
LTFD			DOVE								

Table 2 shows standard deviation and ‘rarity’ of feature, based on VQ data (Stevens and French, 2012), 22 speakers (VQ Whispery voice).

**Appendix B: Table 2:** Example VQ Data, Stevens and French (2013) and analysis

7	91	22	90	74	7	% of speakers with VQ setting		
Lax larynx	Creaky	Whispery	Breathy	Harsh	Tremor	Speaker No	Total	MFCC Classification
2		2	1			001	4	
		2			2	003	2	DOVE
	1			2		009	5	
	1	3	1		1	013	4	
1			1			016	0	
		2			1	018	-2	PHANTOM
	1		1			021	5	
		2				024	-4	PHANTOM
	1	2	1	2		027	9	
	1			2		033	5	PHANTOM
		3	1			035	-1	
	2					037	5	PHANTOM
	1					039	0	
		2	1			040	5	PHANTOM
		2				058	-5	PHANTOM
					2	063	5	WORM
			1			077	-2	PHANTOM
			1		1	086	-1	
	1		1			093	1	PHANTOM
						099	-2	
			1			103	0	PHANTOM
			1			107	2	WORM

Analysis of results revealed several speakers who had similar VQ profiles and zoo position. Those speakers were not necessarily classified in one of the four quartiles, but occupied adjacent points or clustered. Speakers 118 and 009 (LTFD) had extremely similar VQ properties particularly for features that were considered relatively distinct across the group. These were Pharyngeal constriction (23% of speakers), raised larynx (22% of speakers), tense larynx (36% of speakers) and nasal (22% of speakers). Speakers 033 and 051 (LTFD) had similar tense larynx scores. Speakers 026 and 037 (LTFD) had very similar VQ, with features considered highly distinct across the group. They were the only two speakers with VQ close jaw (2% of speakers) and also only two of four speakers scored against minimised range (4%). Speakers 026 and 037 also scored high at +4 or +5 for creaky voice (8% of speakers). Whilst not assigned animal classifications, these speakers occupied adjacent space in the normal position. A tight cluster of x3 speakers in normal was noted 113, 078, 034 (LTFD engine). However, no explanations were found in terms of VQ, SNR or net speech. Interestingly, these speakers were also found in close proximity on the MFCC zoo plot. Speakers 031 and 002 appeared similar to the ASR (both MFCC and LTFD engine, normal classification). No conclusive explanation was found - although both had quite similar VQ scores, particularly for raised larynx (22% of speakers).

### **LTFD Summary**

Of the 100 speakers assessed, 22 speakers had the voice quality whispery. Of those, 9 were most likely to produce FR by the ASR (LTFD engine). Further research would be required to test the hypothesis that poor performers (as per the MFCC engine) or those more likely to produce false rejects (LTFD) have a tendency to produce higher scores for phonatory VQs. For example, an alternative hypothesis might be that the normative data (UBM) is not correctly balanced in terms of those features, although that would still suggest a weak link between certain VQ criteria and zoo plot position.

### **MFCC Summary**

Of the 22% of DyViS speakers were scored in the VQ category whispery with 9 of those 22 speakers classified as Phantoms, accounting for 90% of MFCC engine Phantoms. The majority of worms (2 out of 3) were marked as having the VQ whispery voice. Conversely only 1 out of 8 Doves were marked as whispery and 0 chameleons. This could suggest a possible y-axis divide around this VQ feature however more research would be required. Of all DyViS speakers, 91% had a score for creaky voice. Of the 9 speakers that did not have a score for creaky voice, 7 gained zoo plot classifications. 5 of these were classified as MFCC phantoms with 2 x normal, 1 x dove & 1 x chameleon. Ten speakers did not have a score for VQ breathy voice. Whilst strong correlation was initially not found in terms of classification (2 x doves and 8 x normal) – some clustering was observed in terms of zoo plot position, with 5 speakers forming a close pattern. Clustering was also noticed for those speakers not having the VQ breathy voice (MFCC engine). However, clustering is neither tight nor consistent and other factors may account for zoo plot positioning.

Using Tables for analysis, speakers classified as doves initially appeared to have a greater number of VQ identifiers per speaker (9, 10, 11, 12, 13 features compared to 4, 5, 6, 7, 8 features). However, analysis using zoo plots found a distinct lack of clustering and there was little evidence to support the hypothesis further.

Tense larynx was a VQ feature present for 36% of speakers. All 5 x doves appeared in the subset tense larynx. However, as the tense larynx group is quite large, correlation is not conclusive. Zoo analysis revealed a distinct lack of clustering and the tense larynx group, as a whole, were evenly distributed. However, for the 7 speakers that scored for the VQ lax larynx (normal classification for all instances) a distinct cluster was observed towards the mid left side of the zoo plot.

The raising or lowering of the larynx can influence pitch variation for F1 and F2 (Nolan and Grigoras, 2005). Others have noted links between formant measurements and the position of the larynx (Gold, French, Jessen et al., 2013). As F1, F2 and F3 are measured by the ASR LTFD engine this suggests causality for zoo positioning in terms of VQ. The results require further examination, but this could



demonstrate commonalities in terms of zoo plot placement for certain laryngeal settings when using the LTFD engine.

For 10 of the VQ criteria all 100 speakers scored zero. For an additional 6 criteria only 4% (or less) of the speakers scored greater than zero. Conversely, 2 features were present for almost all speakers. Analysis, in terms of zoo positioning, focused on VQ data for which groups of speakers could be compared well against the majority. It is conceded that constraints in the variability of the data meant that not all VQ features could be analysed for zoo positioning.

Initially and at a high level, links between ASR classification and VQ were not observed with random distribution. Those speakers with distinct voices from VQ analysis did not directly correspond with doves and speakers regarded as not distinctive, by VQ, did not directly correspond with worms. Speakers with similar VQ profile were not often adjacent, although there were examples where they were. Results were partially to be expected, since VQ is a subjective measurement using different metrics to MFCC or LTFD extraction. However, further analysis of zoo plot positioning suggested some support to the hypothesis that VQ is linked to zoo plot position and ASR performance in some instances. Ongoing research is suggested to examine the following:

- i. A suggested link between phonatory features and zoo plot positioning, but not necessarily classification (MFCC engine);
- ii. A suggested link between supralaryngeal features and zoo plot positioning, but not necessarily classification (more prominent in LTFD mode);
- iii. Other VQ features were examined for both engines and were regarded as having a low likelihood of influence on zoo positioning.

# Appendix C

---

## **Preliminary test results, VQ and ASR zoo plot analysis**

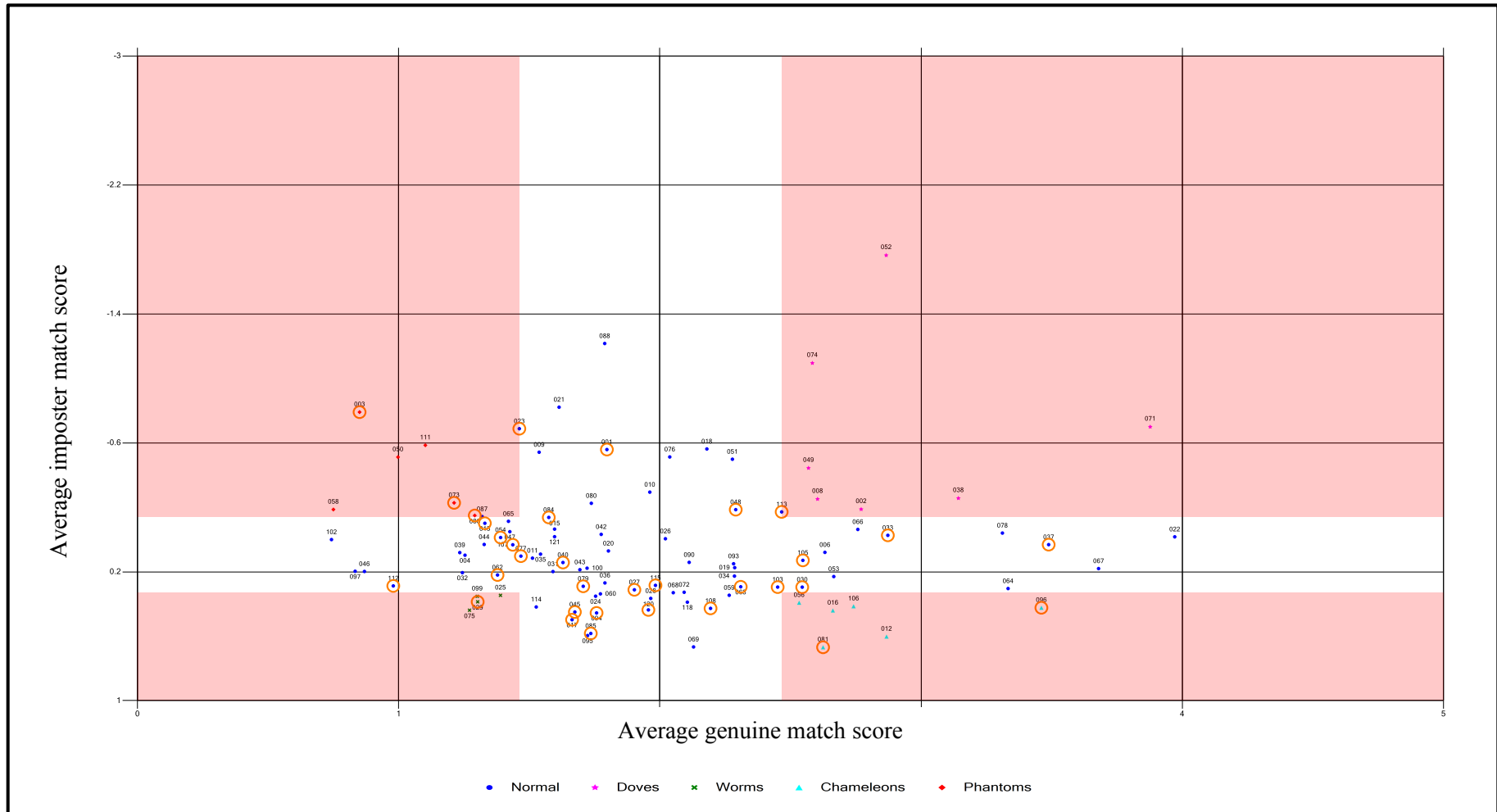
Are there correlations in terms of voice quality from both the MFCC and LTFD engine classifications?

High-level correlations could not be found between VQ and ASR classifications. Although some correlations were noted between certain aspects of voice quality and the zoo plot results for both engines, particularly in the classifications phantoms (MFCC) and chameleons (LTFD). In terms of zoo plot positioning, clusters appeared to highlight speakers with certain supralaryngeal VQ features (LTFD engine). Speakers with certain phonatory VQ features also appeared to cluster (MFCC engine). Further research is required to rule out other factors.

Are there any other clusters or patterns that could indicate commonalities, particularly in terms of voice quality properties?

High-level commonalities could not be found, but in some instances certain speakers with similar voice quality characteristics appeared adjacent or in close proximity on the zoo plot. A tentative link was observed between the MFCC engine and speakers that had certain phonatory features and the LTFD engine and speakers that had certain supralaryngeal features. Zoo plots are a useful tool for examining overall ASR health and can assist in identify outlying speakers that can produce comparison issues. Conclusively determining the exact cause of each zoo plot data position, however, is challenging due to the quantity of variables inherent in the data and the combination of factors likely to influence each plot position. Constraints and limitations in zoo plot interpretation must be recognised.

In the section below 19x zoo plots are presented which informed the examination of VQ (Stevens and French, 2013) against position in 6.5.4.

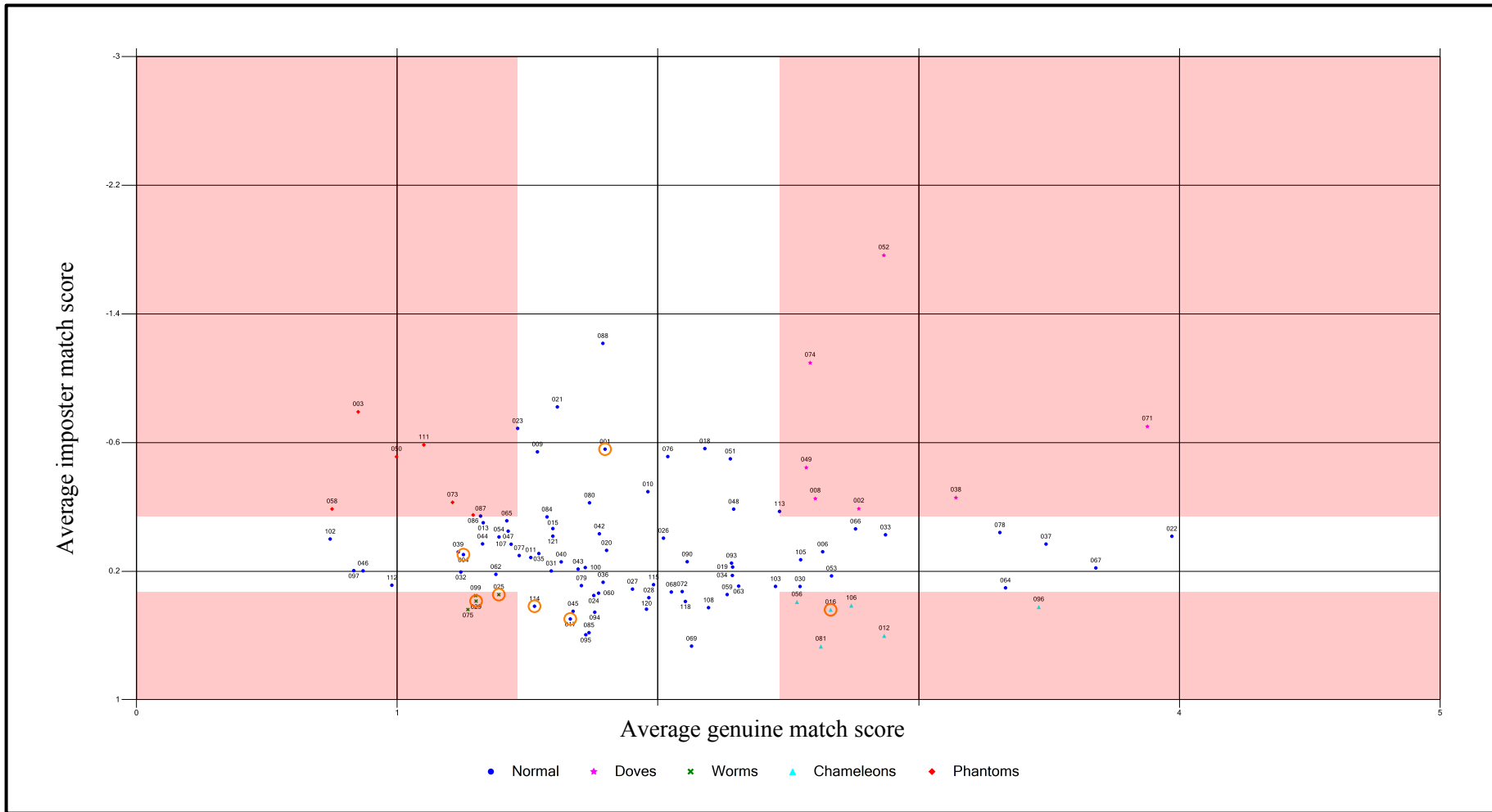


VQ: Speakers with higher sibillance scores [+2 or +3]

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

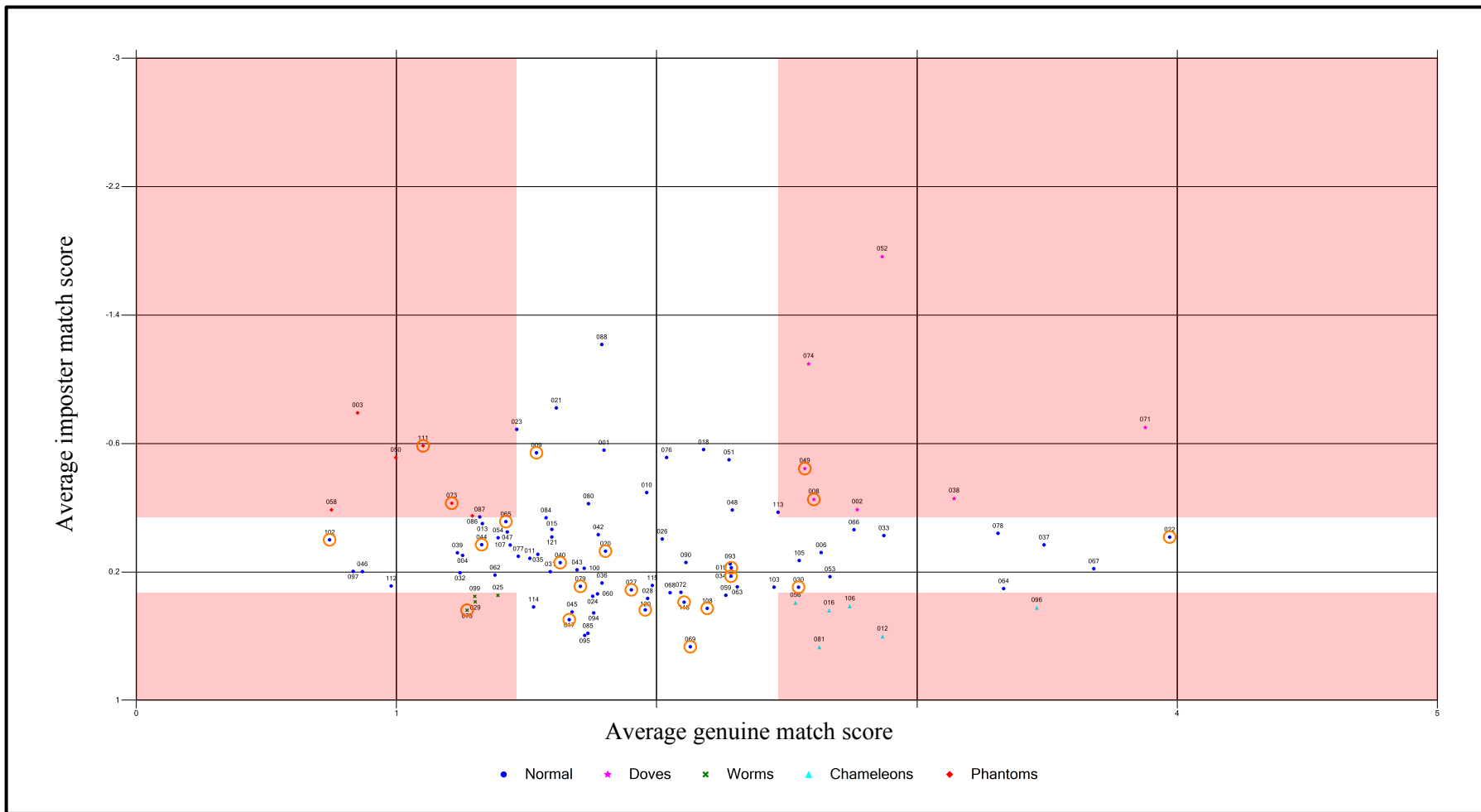
Vocalise 1 ASR: LTFD (scale in LLR)

UBM: CTEST 89 speakers: Interview domain

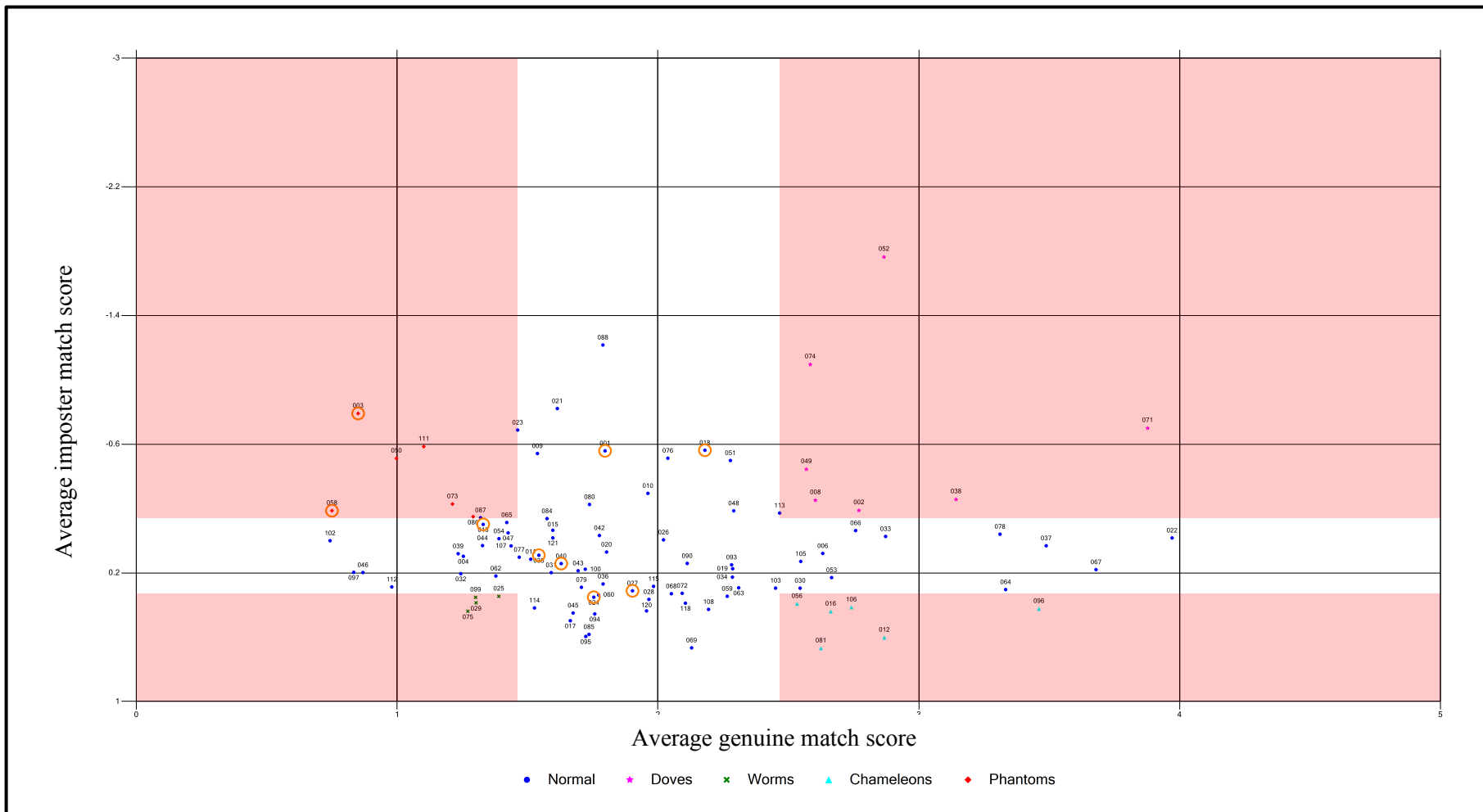


VQ: Speakers with lax layrn scores  
 Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise 1 ASR: LTFD (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain





VQ: Speakers with pharyngeal constriction scores  
 Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise 1 ASR: LTFD (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain



VQ: Speakers with higher whispery scores [+2 or +3]

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: LTFD (scale in LLR)

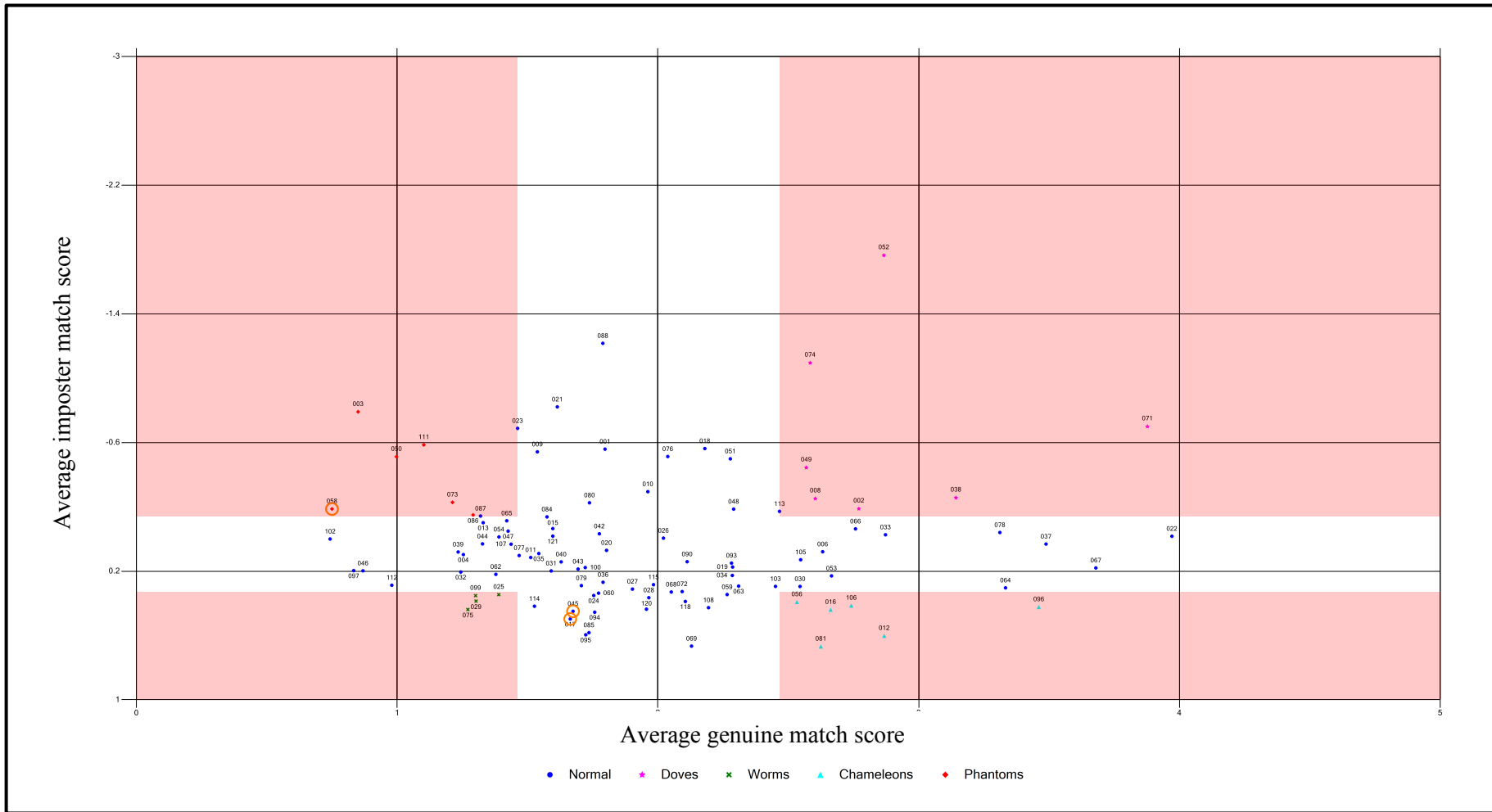
UBM: CTEST 89 speakers: Interview domain









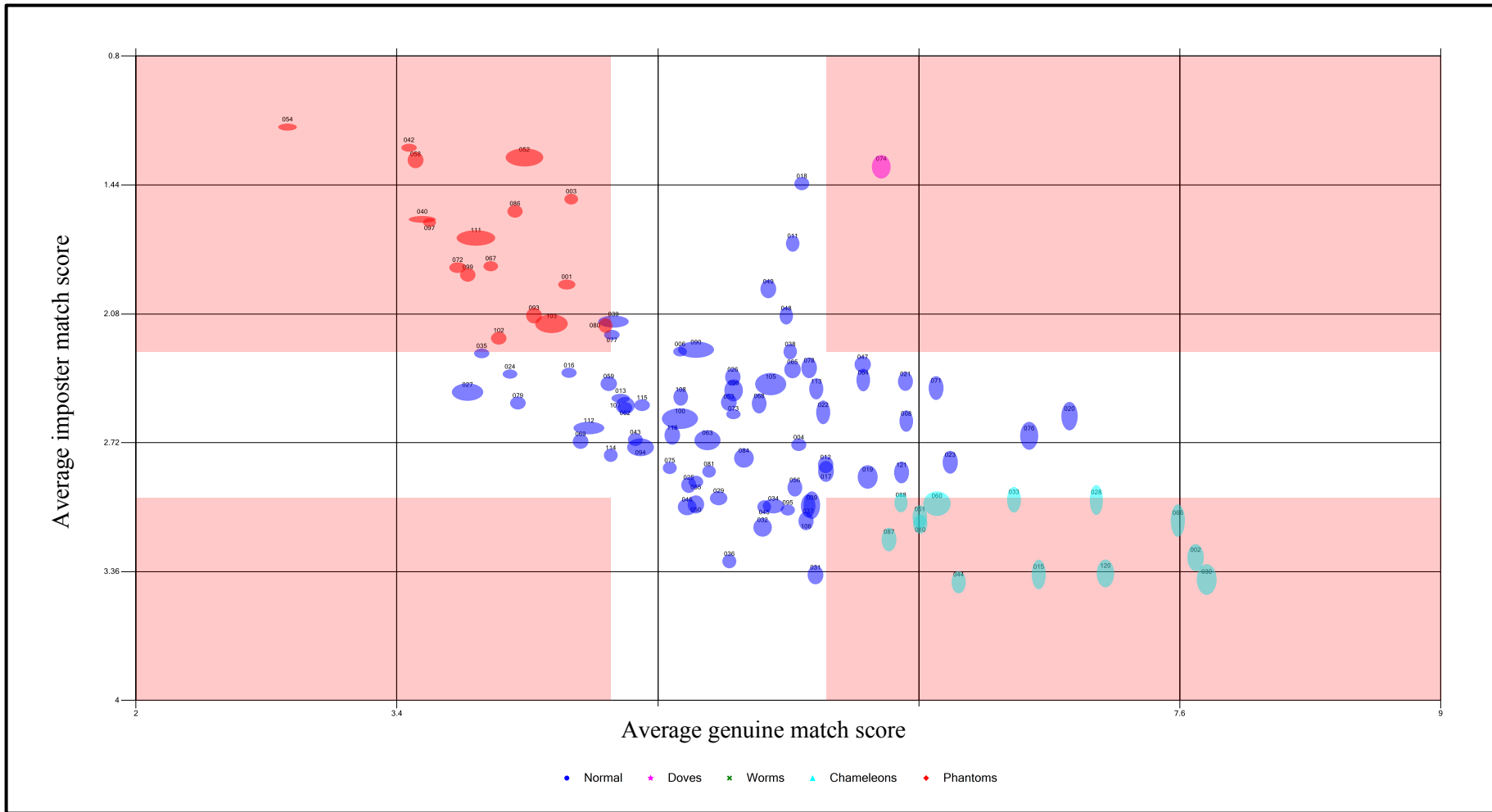


VQ: Speakers without fronted tongue body scores

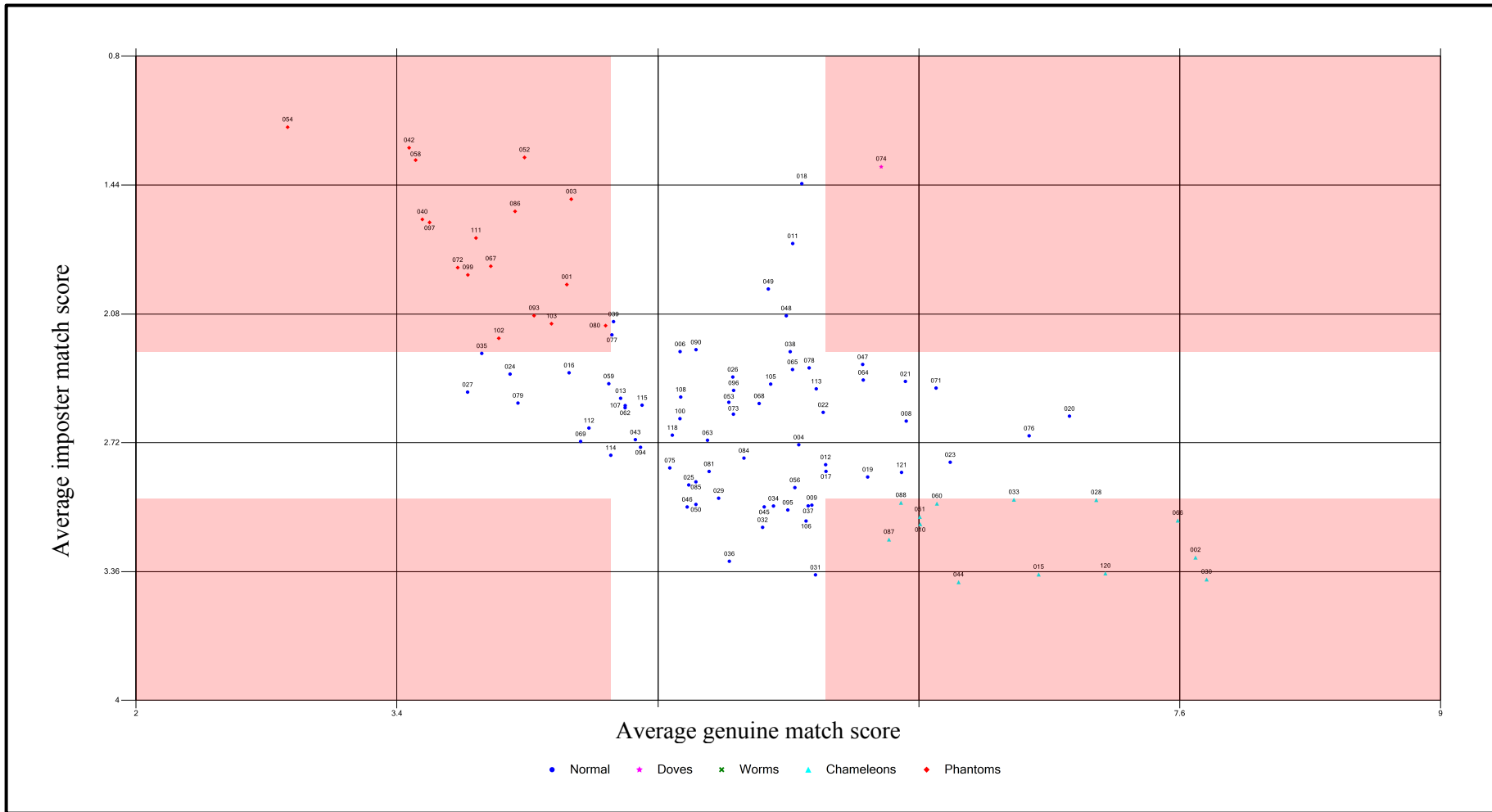
Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: LTFD (scale in LLR)

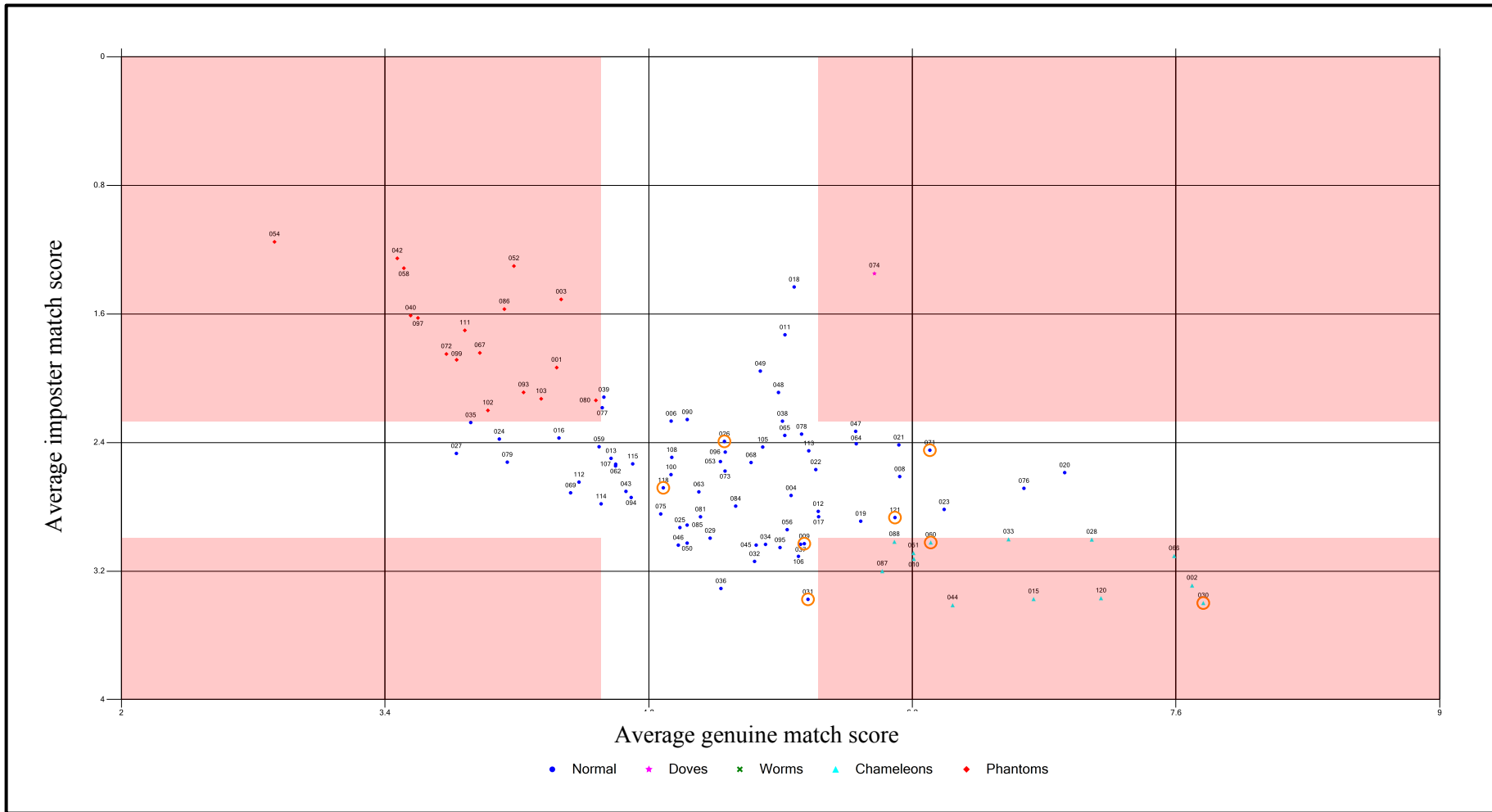
UBM: CTEST 89 speakers: Interview domain



Zoo plot with fat & thin option selected – demonstrating inter and intra variability relative to the total (100) speakers  
 Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain



Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain

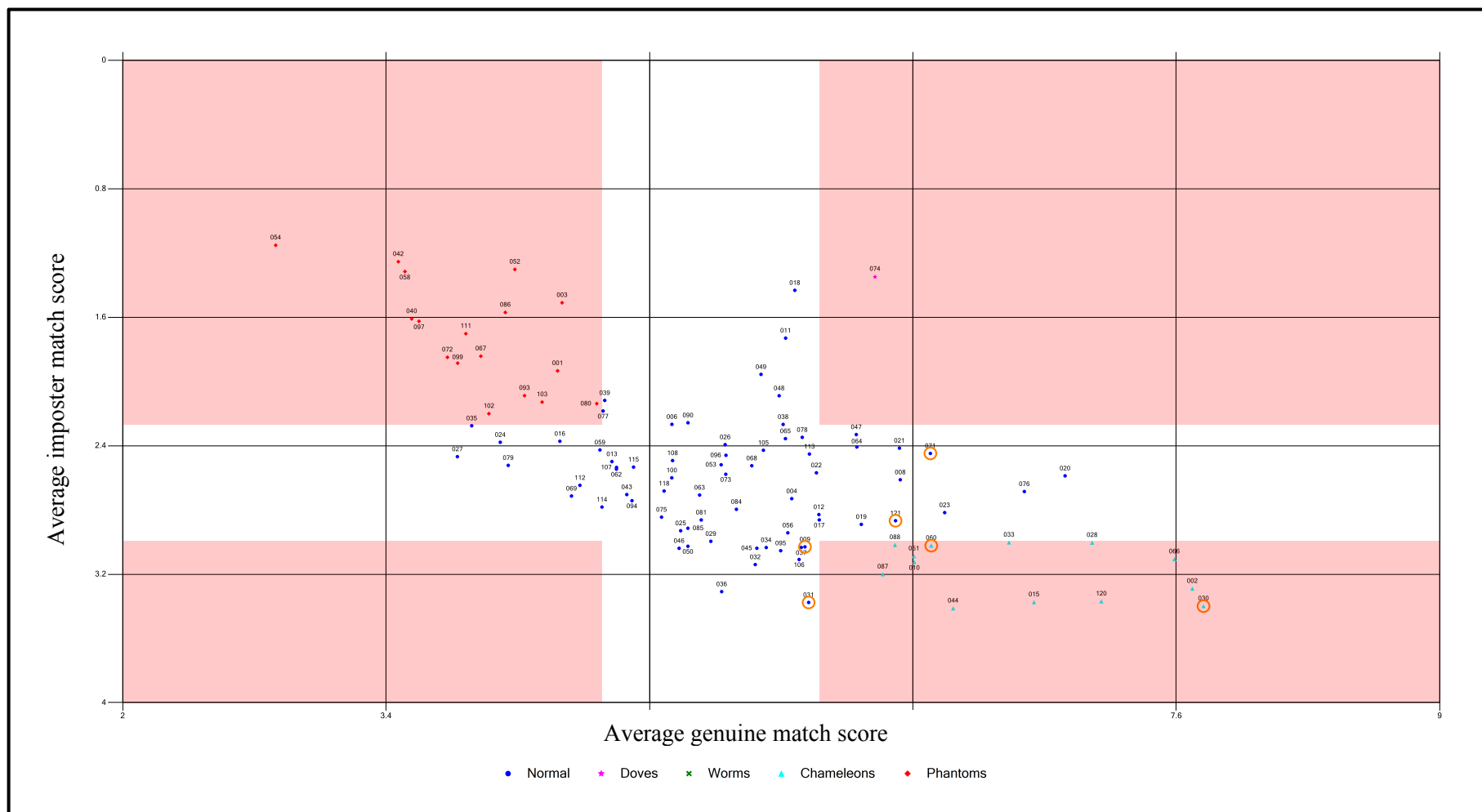


VQ: Speakers with high nasal scores [+3 and +4]

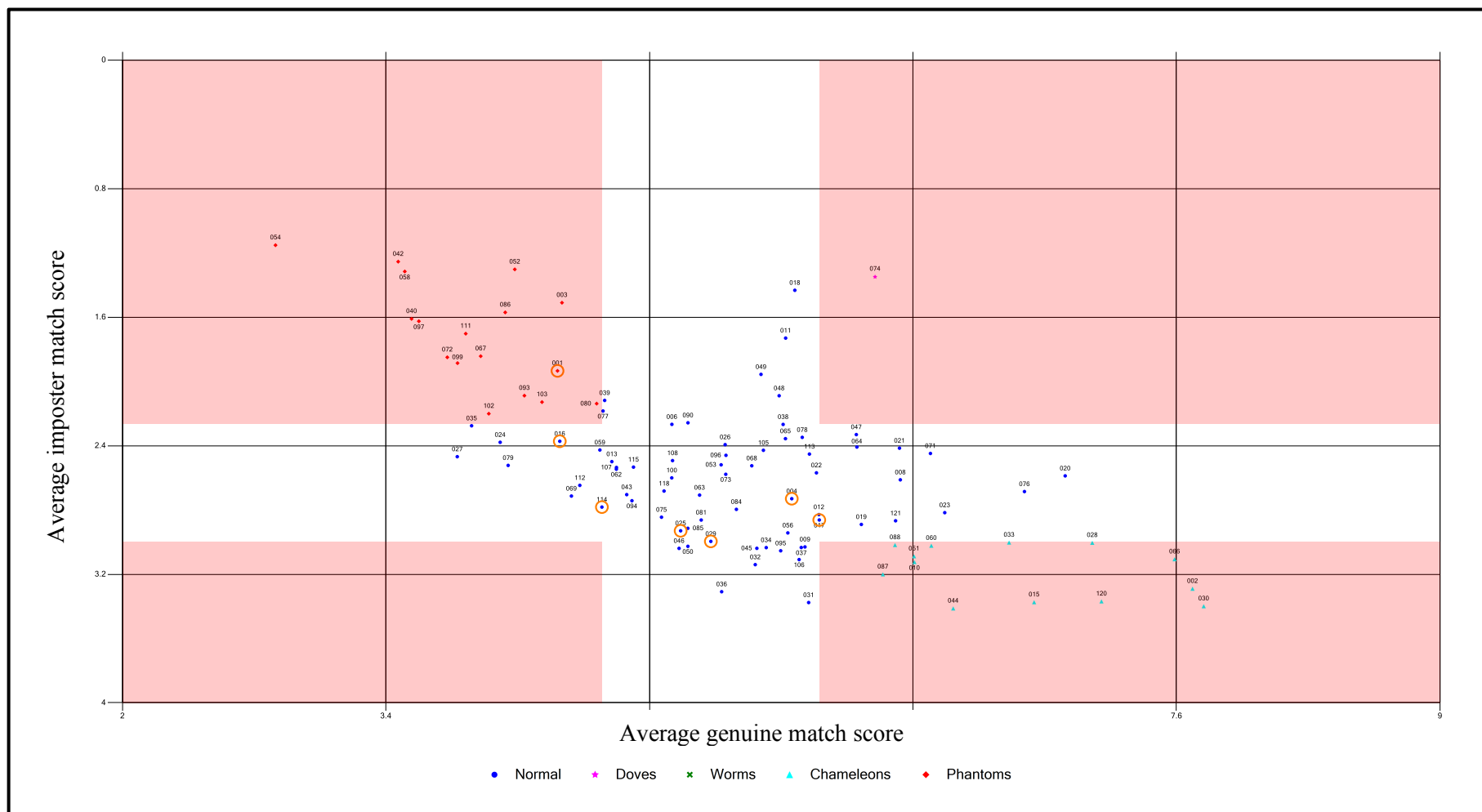
Test material: TypeI DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)

UBM: CTEST 89 speakers: Interview domain

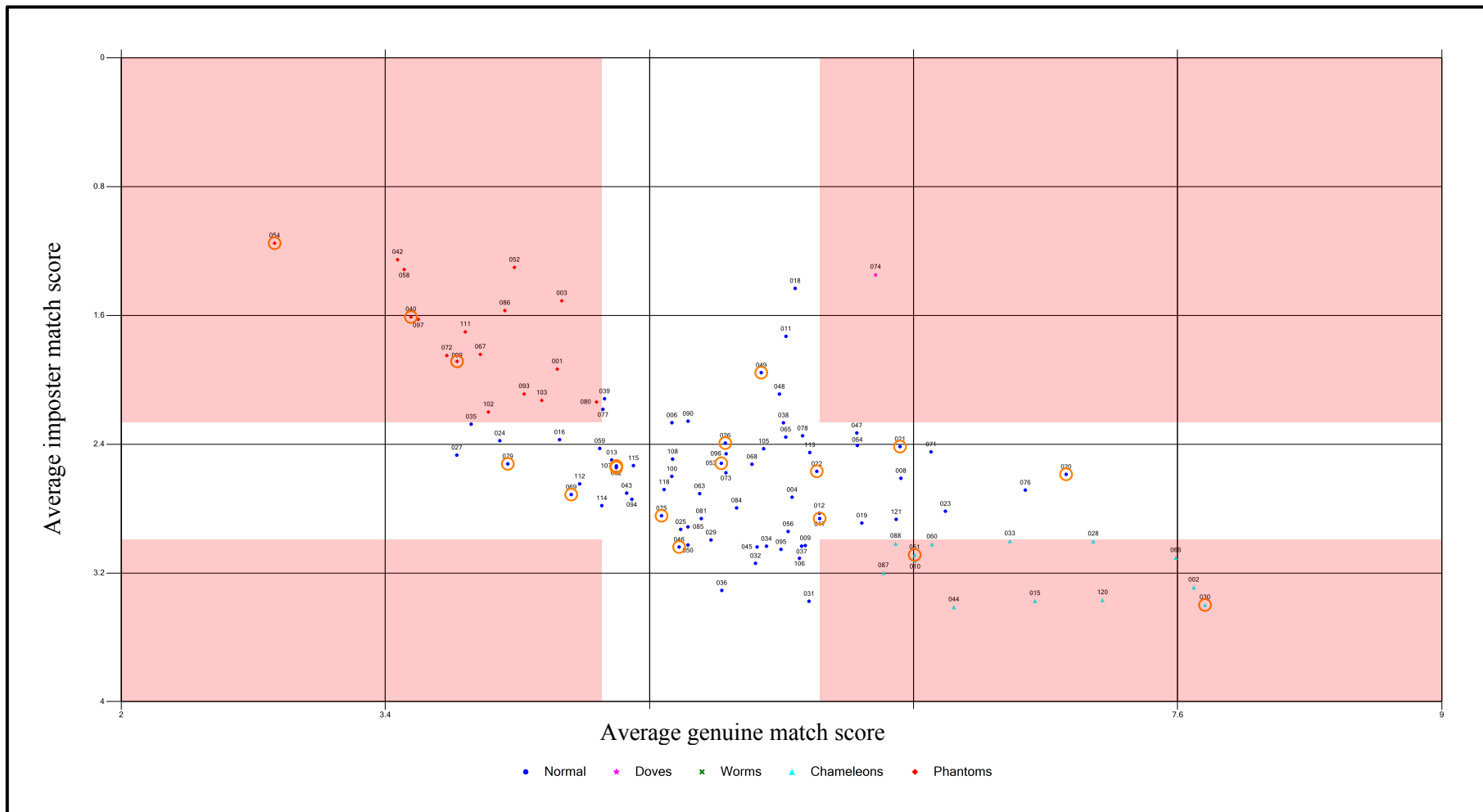


VQ: Speakers with nasal [scores of +3 only]  
 Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain



VQ: Speakers with lax layrn scores  
 Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]  
 Vocalise engine: MFCC GMM-UBM (scale in LLR)  
 UBM: CTEST 89 speakers: Interview domain





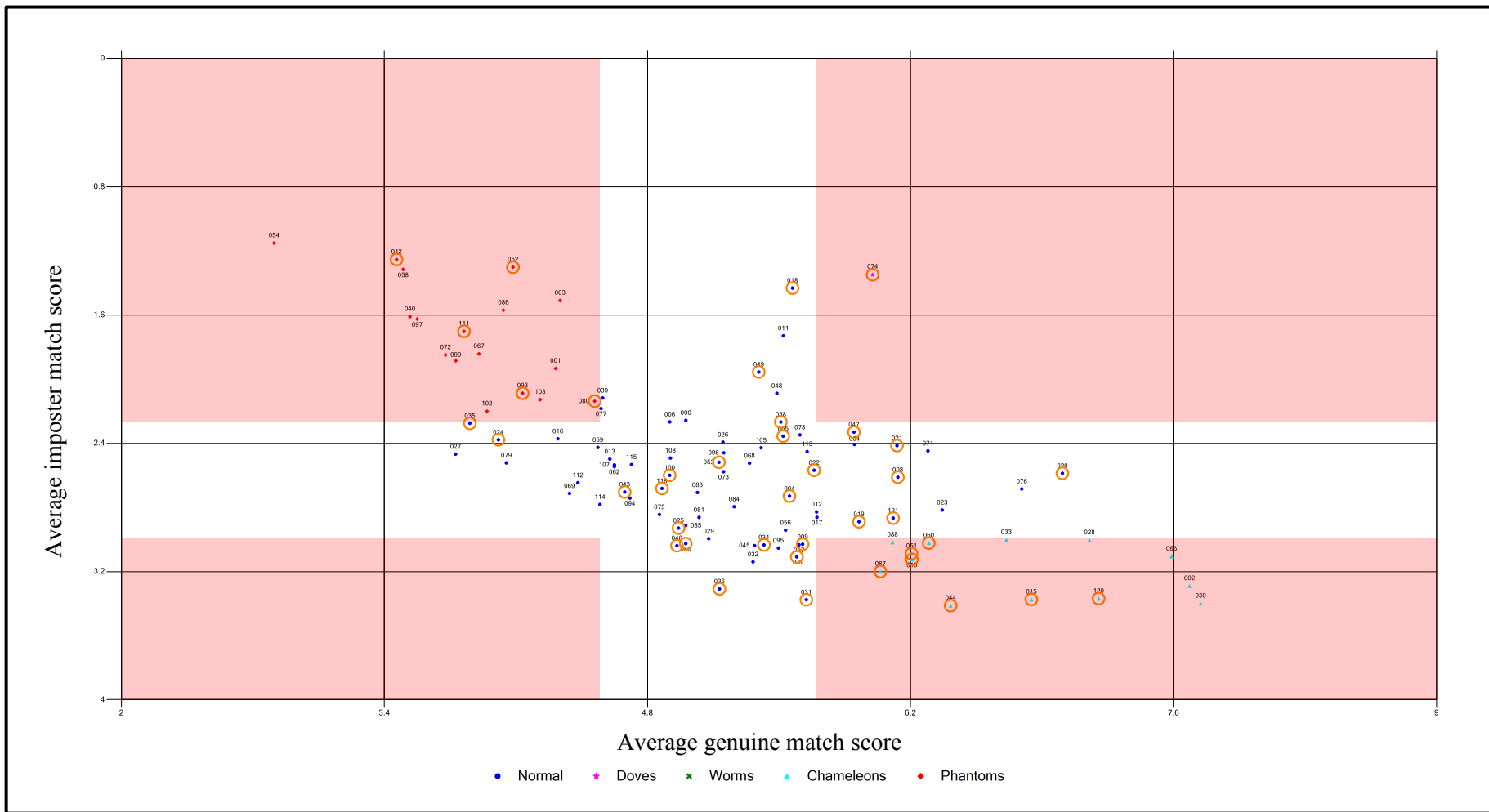
VQ: Speakers with extensive range scores

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)

UBM: CTEST 89 speakers: Interview domain



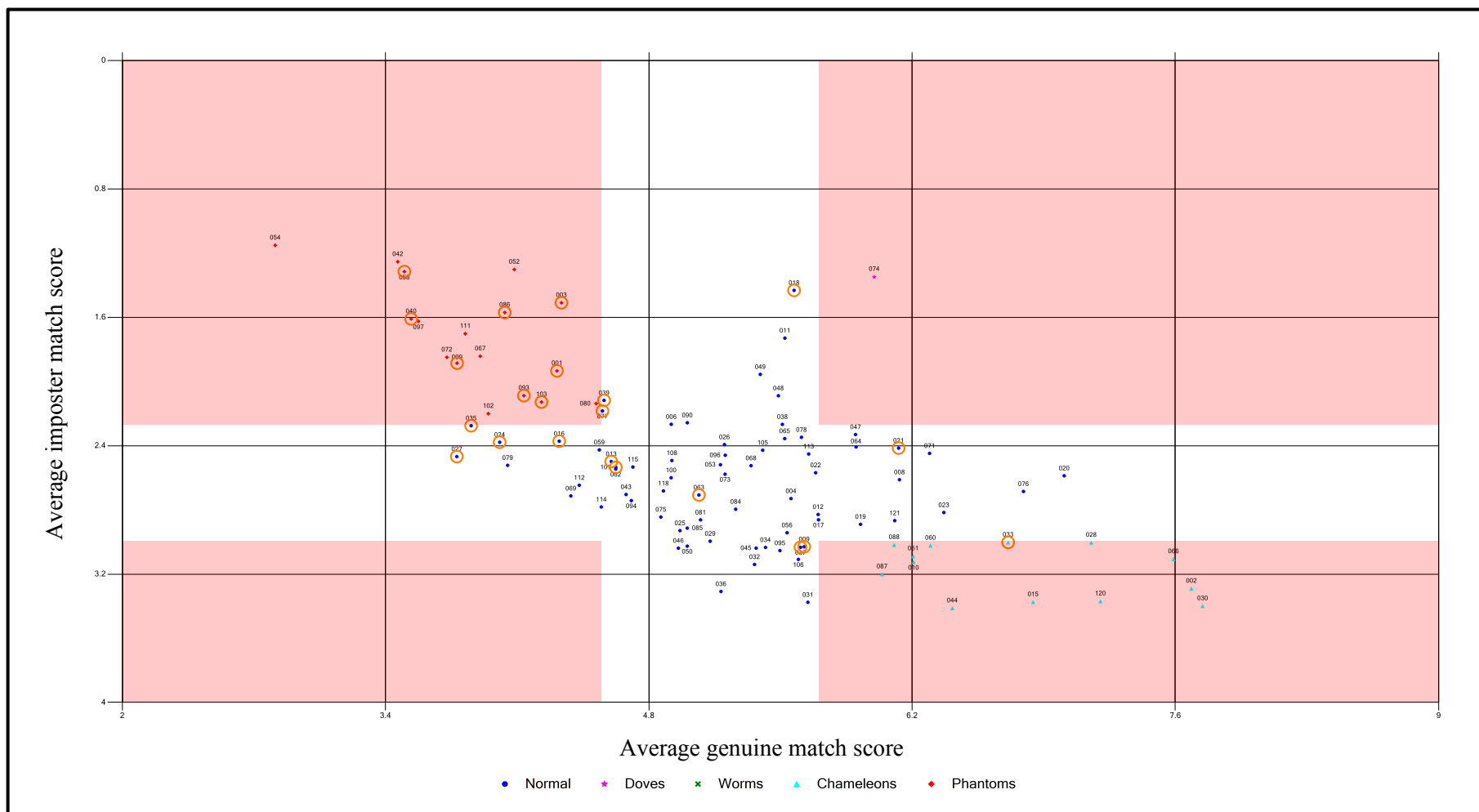


VQ: Speakers with sibilance scores

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)

UBM: CTEST 89 speakers: Interview domain

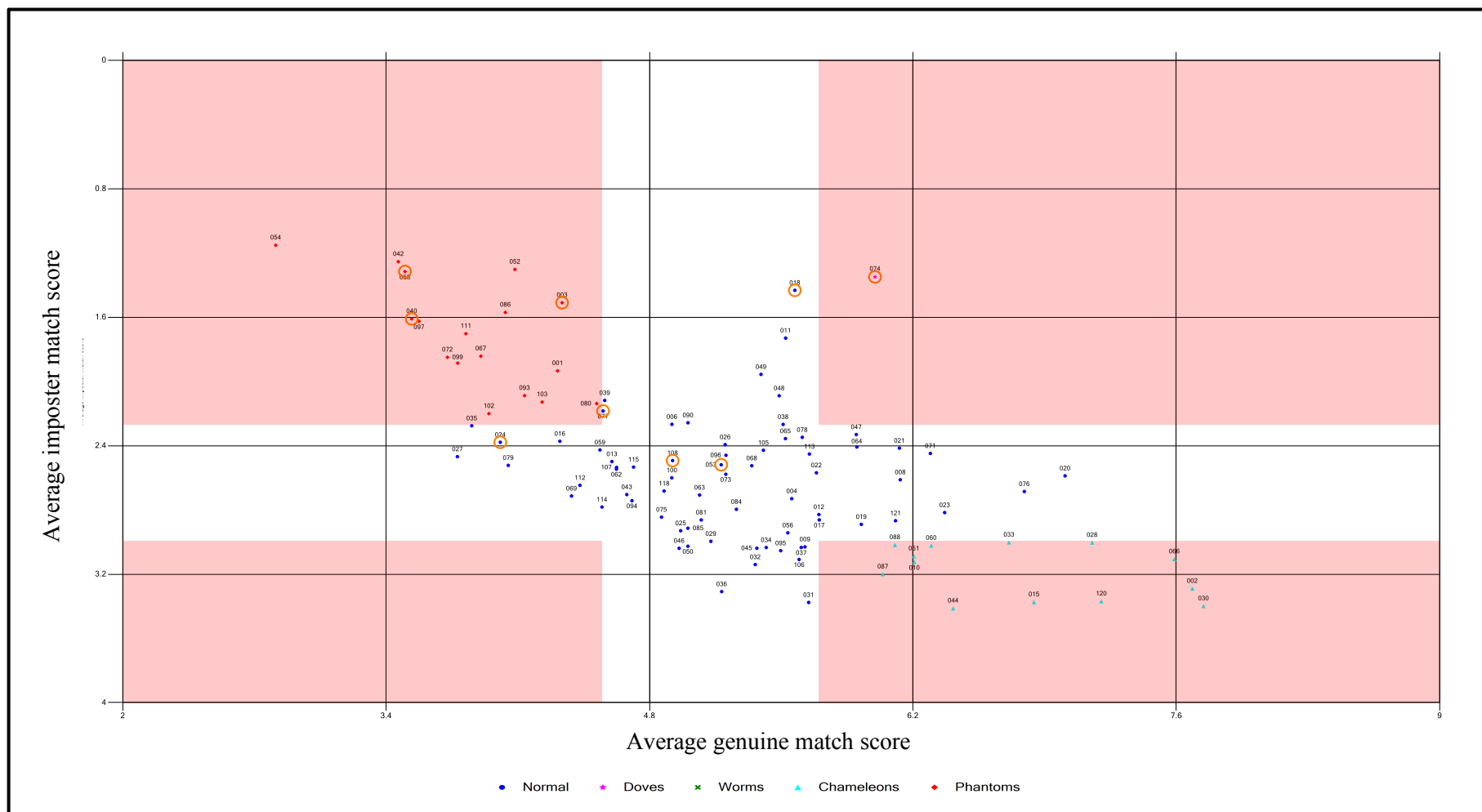


VQ: Speakers with whispery scores [note qty of Phantoms]

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)

UBM: CTEST 89 speakers: Interview domain



VQ: Speakers without creak

Test material: Task I DyVIS 1m SM x 1m + 1m + Residual [29,700 imposter, 300 genuine]

Vocalise 1 ASR: MFCC GMM-UBM (scale in LLR)

UBM: CTEST 89 speakers: Interview domain

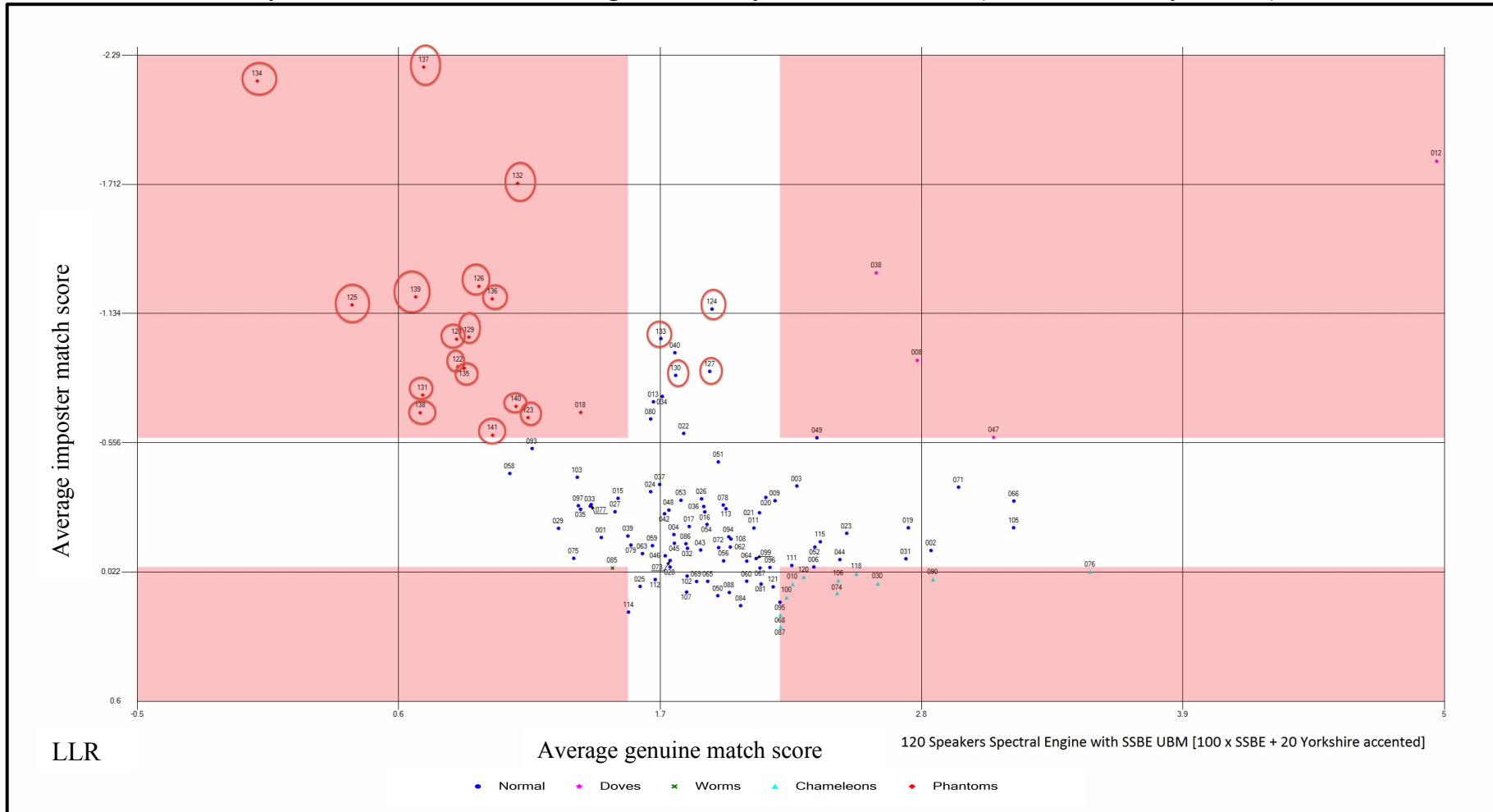
# Appendix D

---

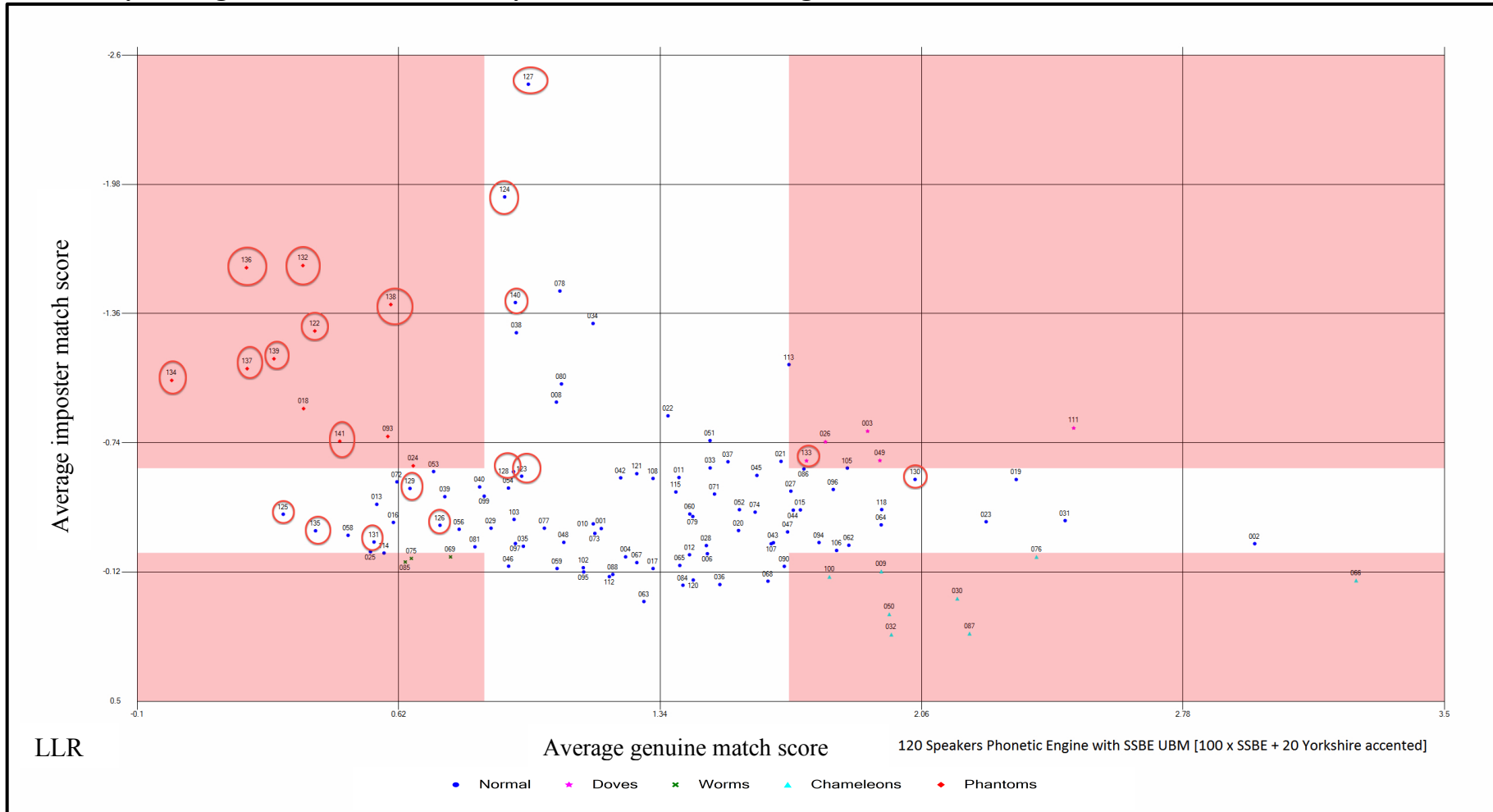
Below, 8x zoo plots are presented from the preliminary experiments. These pertain to experimentation re the addition of Pakistani and Yorkshire accented data to the DyViS SSBE accented data (baseline). They are presented here as a record of the analysis completed and the conclusions reached in chapter 6.

Investigation showed that accent data, different from the SSBE accented data, did indeed cluster, but that no direct correlation between zoo plot position and accent could be determined which couldn't also be explained acoustic/channel differences or by the addition of DyViS to the normative data.

Yorkshire accented + DyVIS : VOCALISE MFCC Engine. Note DyViS Normative set (results artificially skewed)

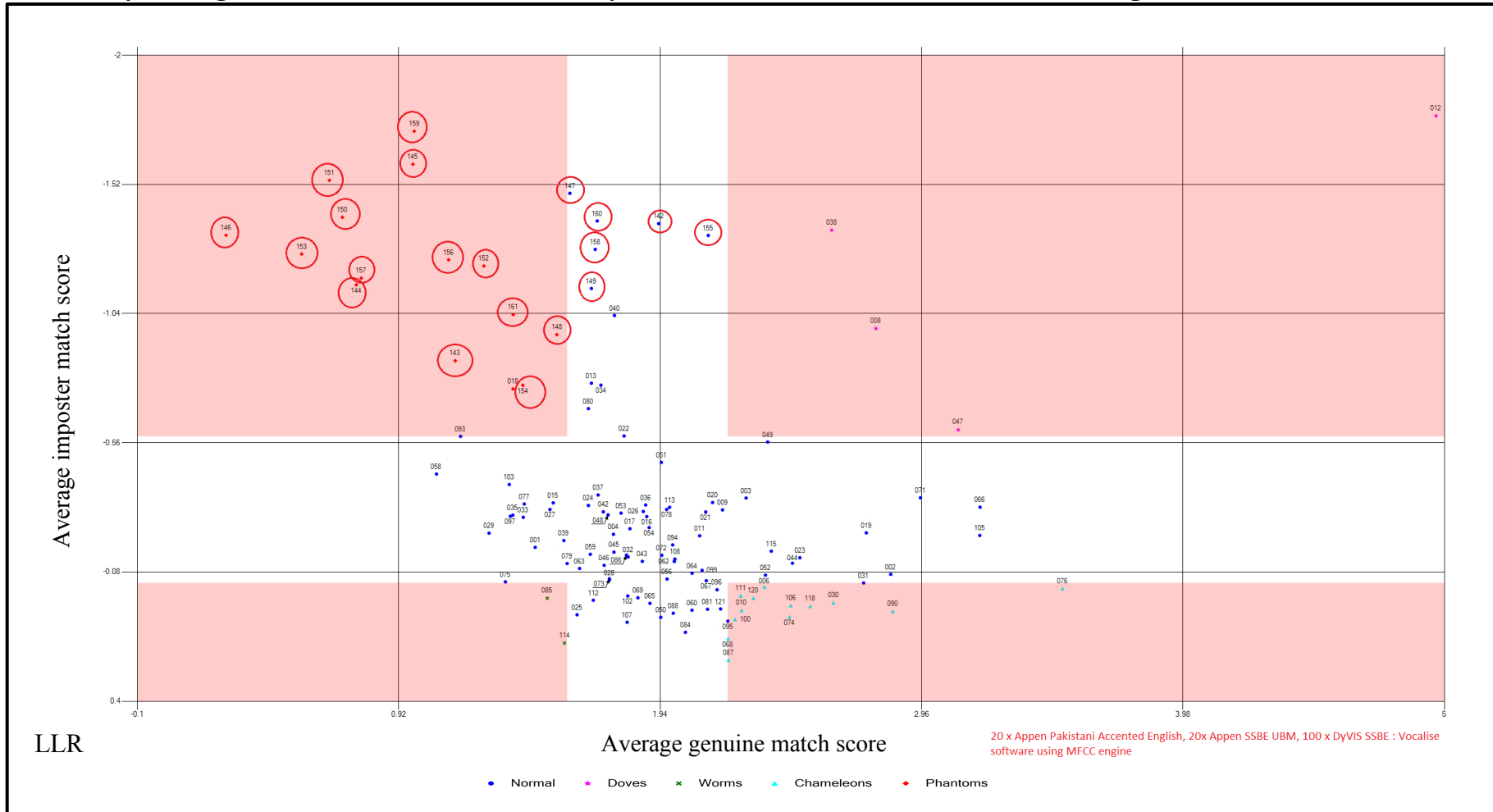


**Preliminary Testing. Yorkshire accented + DyVIS: Vocalise, LTFD Engine**

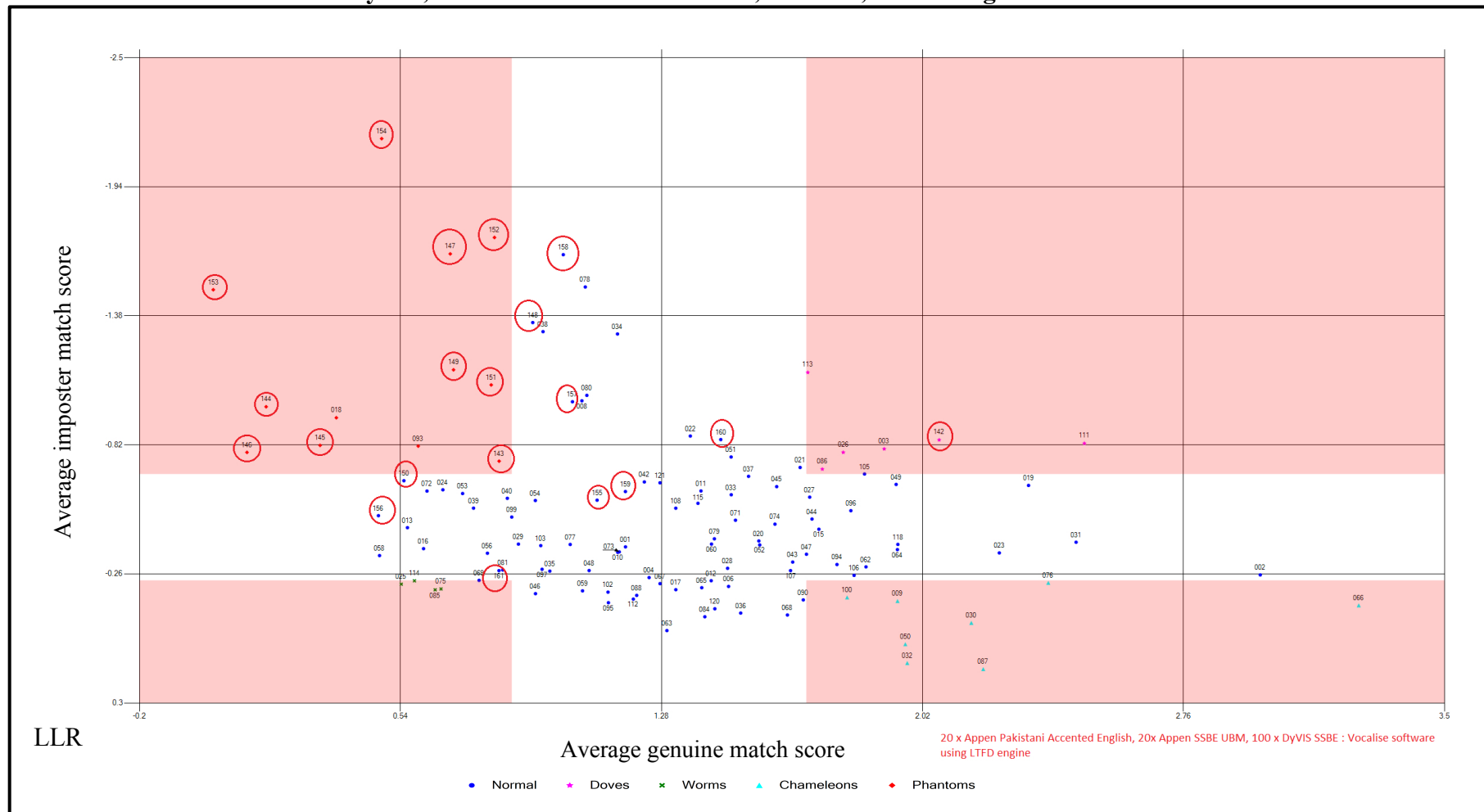




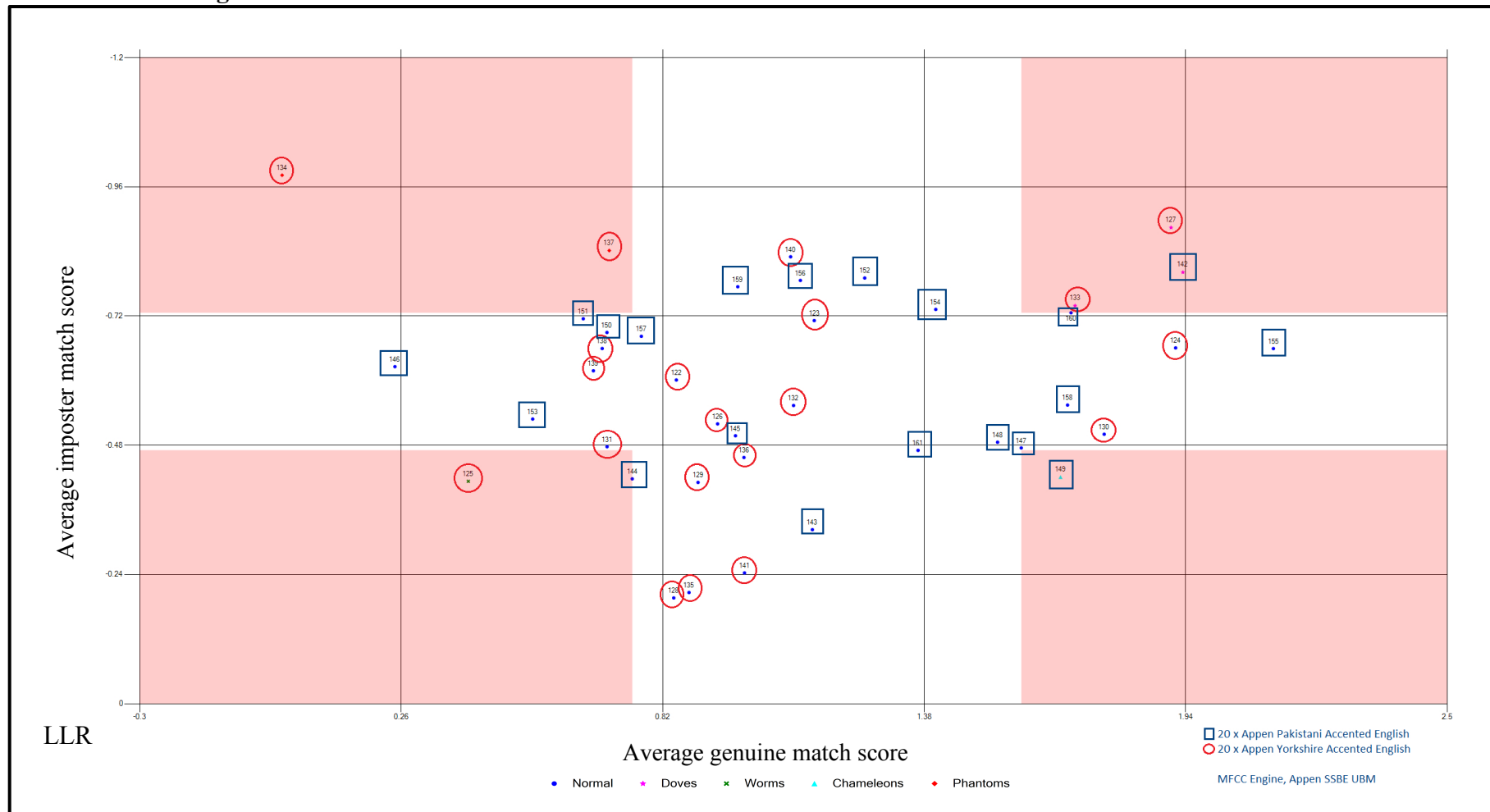
**Preliminary Testing. 20x Pakistani accented + 100x DyVIS SSBE: Vocalise, GMM-UBM, MFCC Engine**

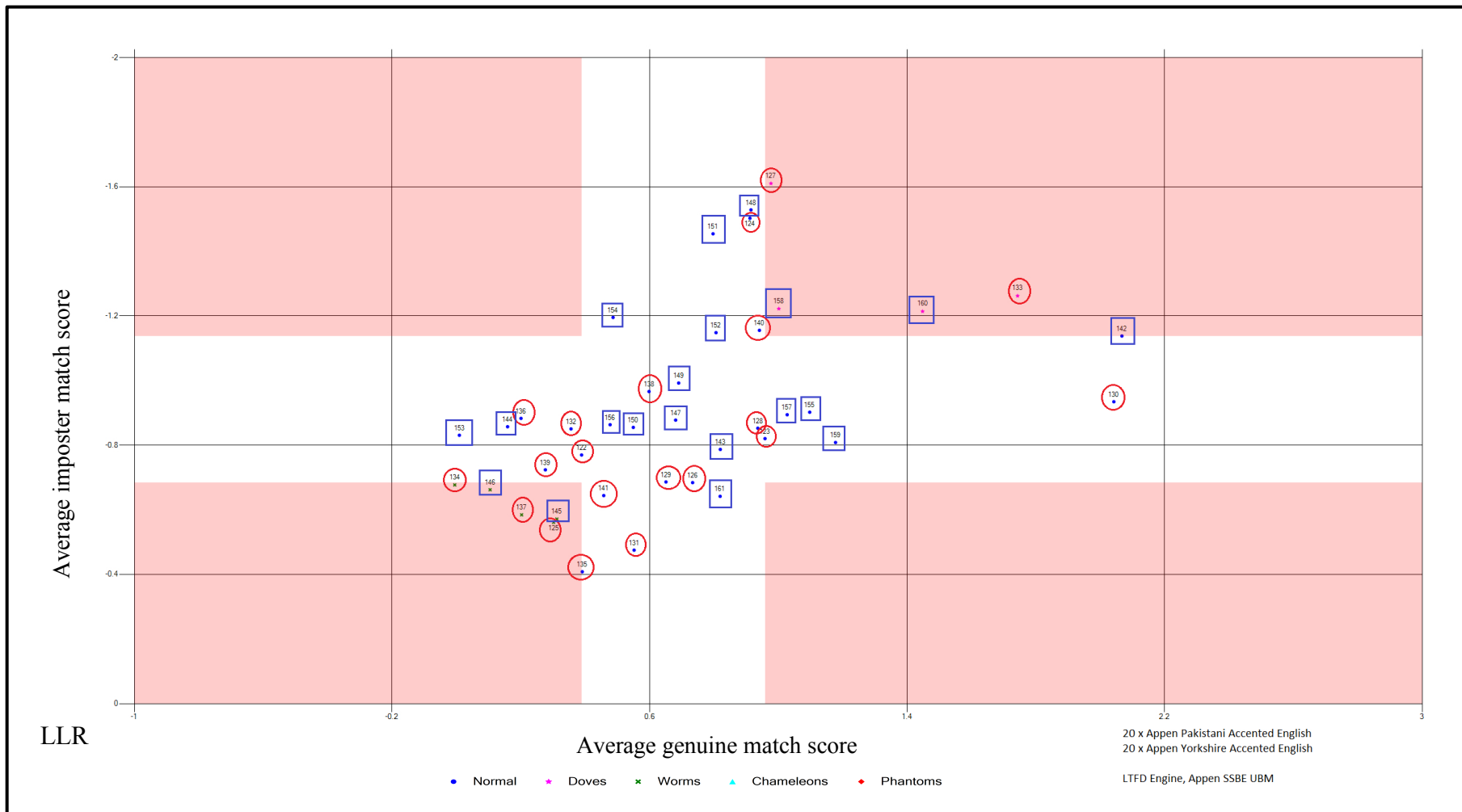


20x Pakistani accented + 100x DyVIS, SSBE: Vocalise GMM-UBM, Vocalise, LTFD Engine.



**Preliminary Testing. 20x Pakistani accented + 20x Yorkshire accented: Vocalise MFCC Engine.**  
**Blue squares = Pakistani accented. Red circle = Yorkshire accented.**  
**Note – no clustering found.**

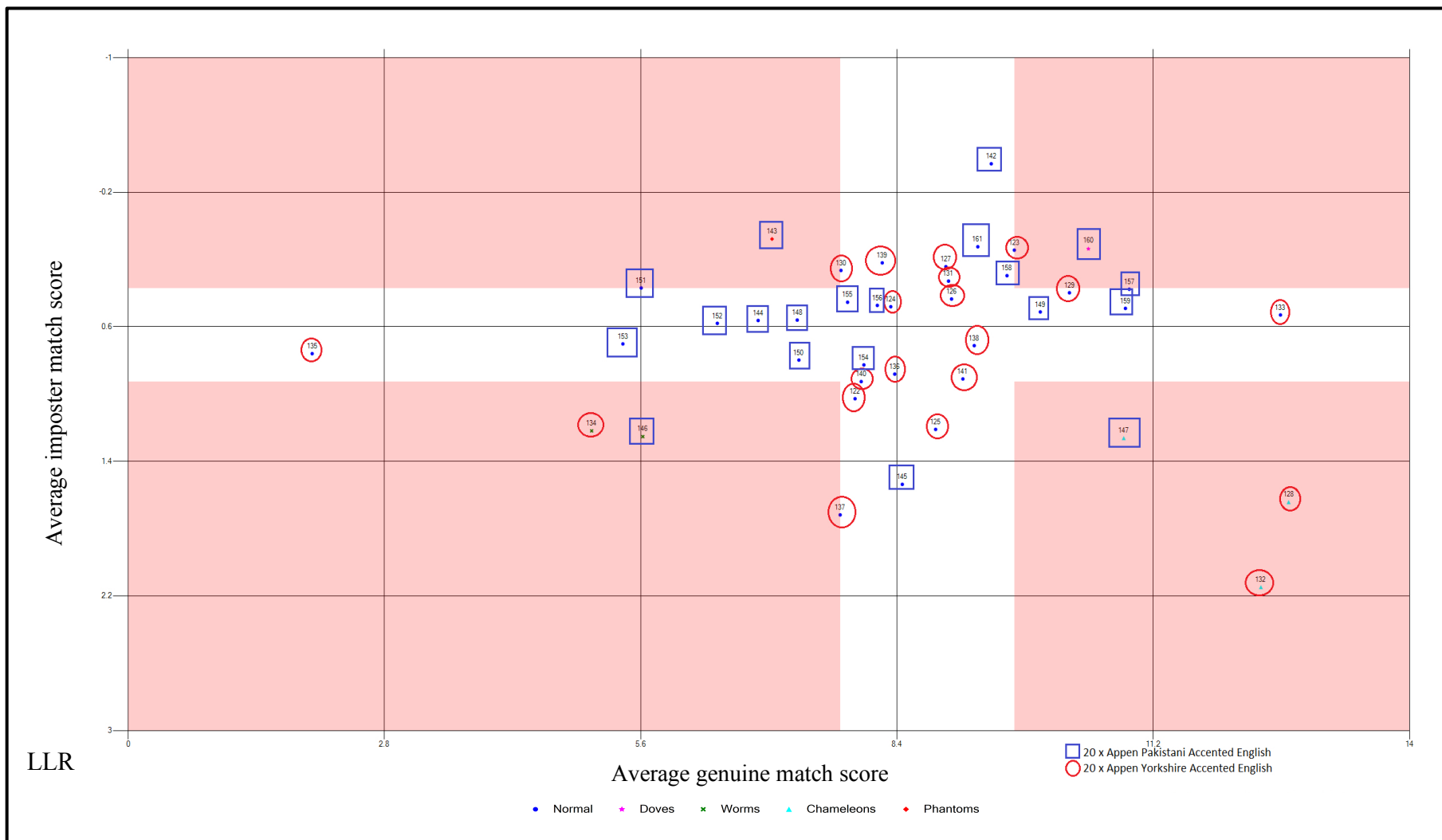




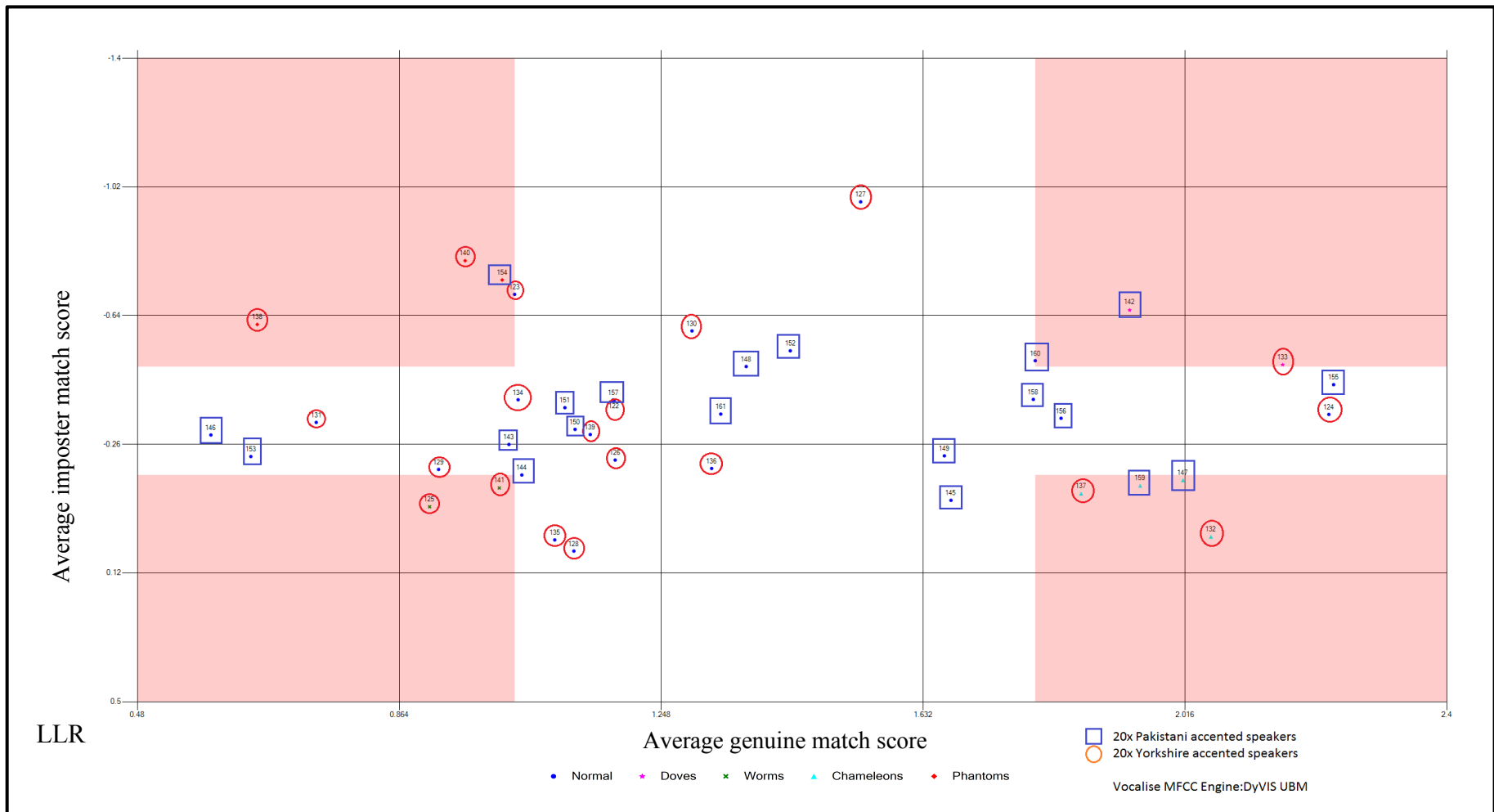
**Preliminary testing, 20x Pakistani accented + 20x Yorkshire accented: Vocalise LTFD**

**Blue squares = Pakistani accented. Red circle = Yorkshire accented**

**Note – no clustering found.**



**Preliminary testing, 20x Pakistani accented + 20x Yorkshire accented: Vocalise, GMM-UBM ASR system. MFCC.**  
**Blue squares = Pakistani accented. Red circle = Yorkshire accented.**  
**Note – no clustering found.**



**Preliminary testing, 20x Pakistani accented + 20x Yorkshire accented: DyVIS UBM. Vocalise, GMM-UBM ASR. MFCC.**

**Blue squares = Pakistani accented. Red circle = Yorkshire accented**

**Note – no clustering found.**

# Appendix E

---

## **.GIF file zoo plots (frames) re frequency bandwidth experiments.**

.GIFs have been generated to display the movement of speaker performance (zoo plots) and the overall system performance (LR plots) – please see additional material section for details. The .gif animations are presented to demonstrate the influence of frequency bandwidth on ASR performance.

The individual frames from the preliminary testing .GIF are presented below as a record. They demonstrate:

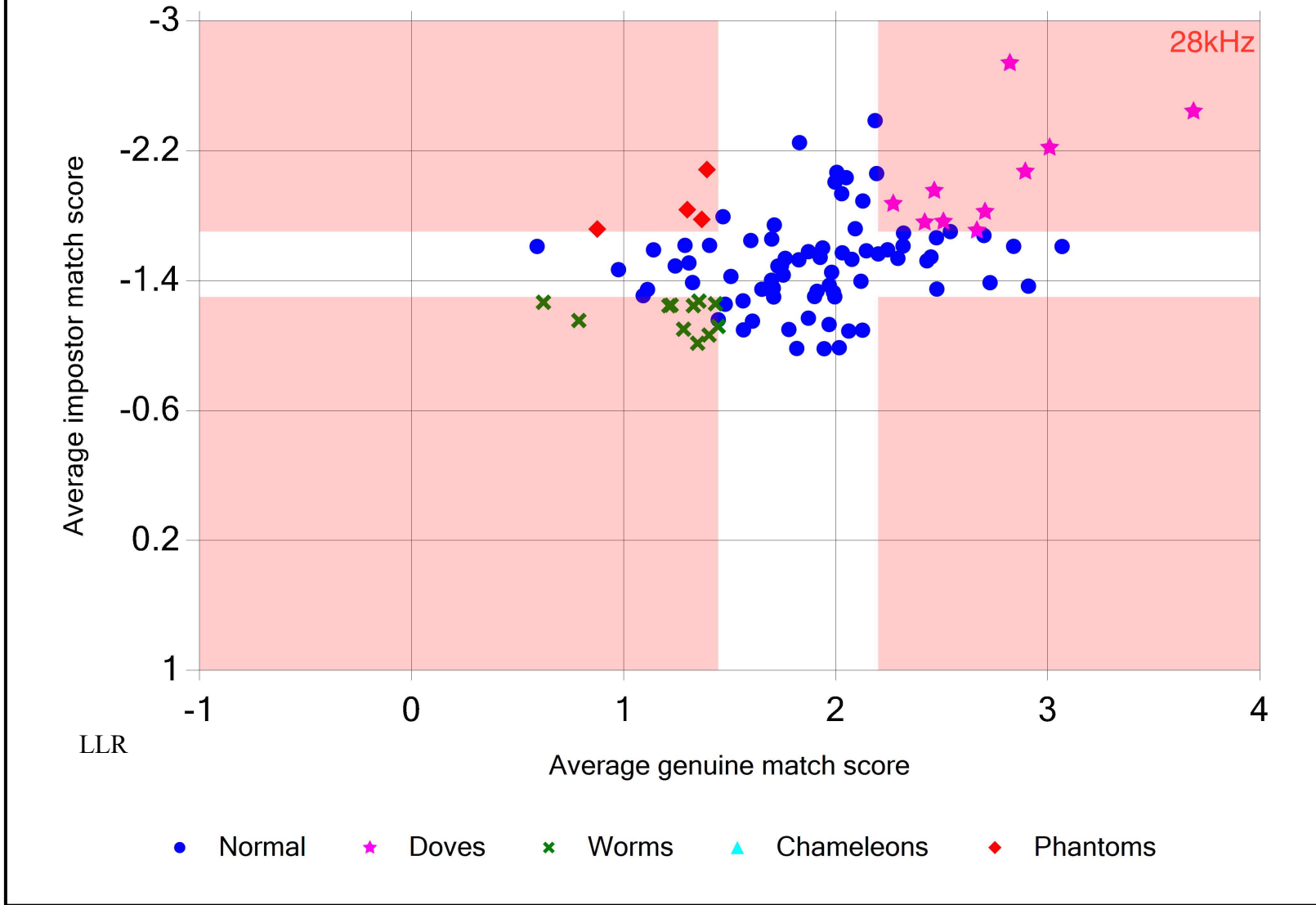
- i. The zoo plot movement of speaker positions from the upper right to lower left position i.e. high TP and low TN transitioning to low TP and high TN as the frequency bandwidth was constrained;
- ii. The LR plot movement (H0 and H1 distribution) as frequency bandwidth was constrained – note the axis scale movement.



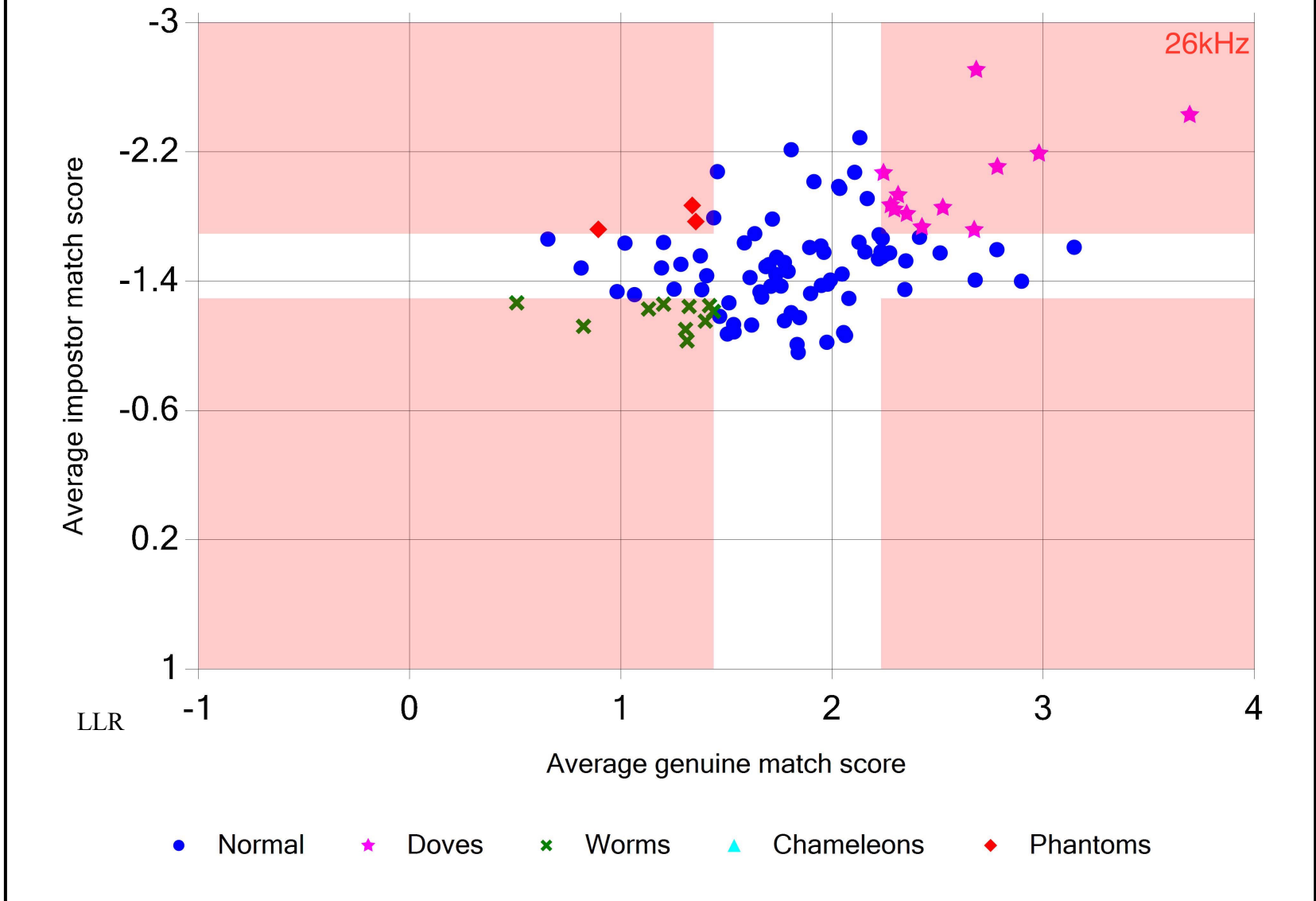




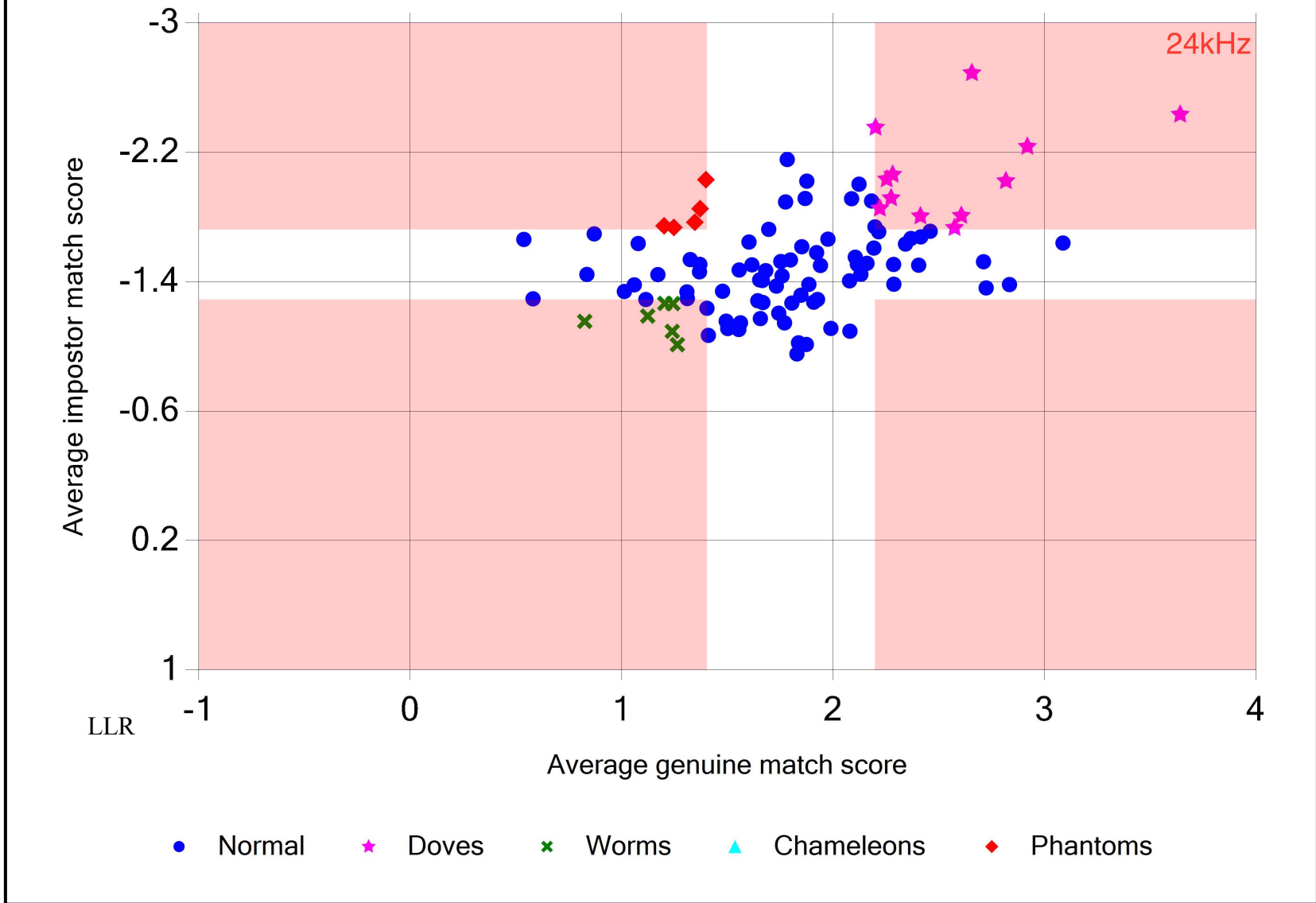
Sample Rate 28kHz, Frequency Bandwidth 0-14kHz

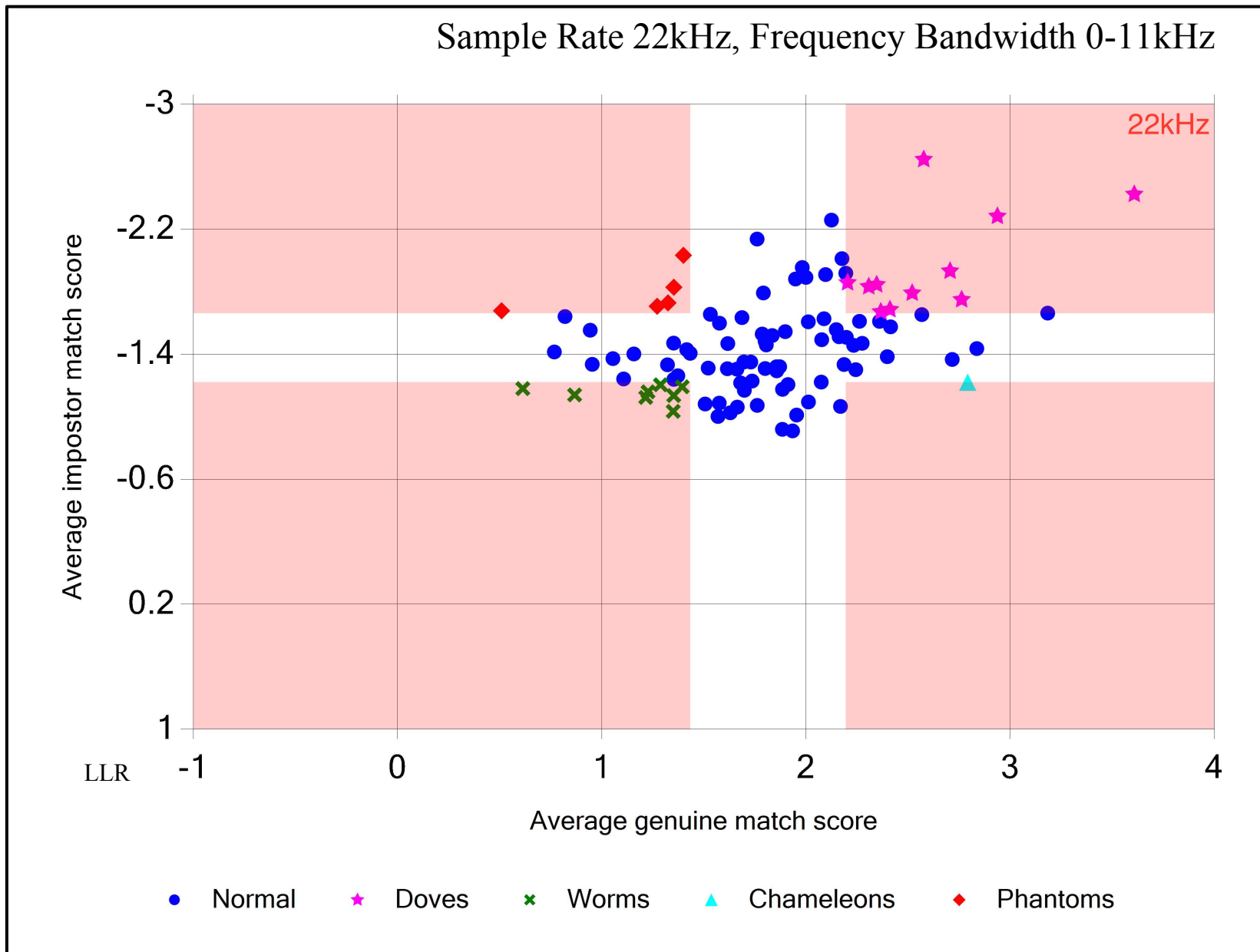


Sample Rate 26kHz, Frequency Bandwidth 0-13kHz

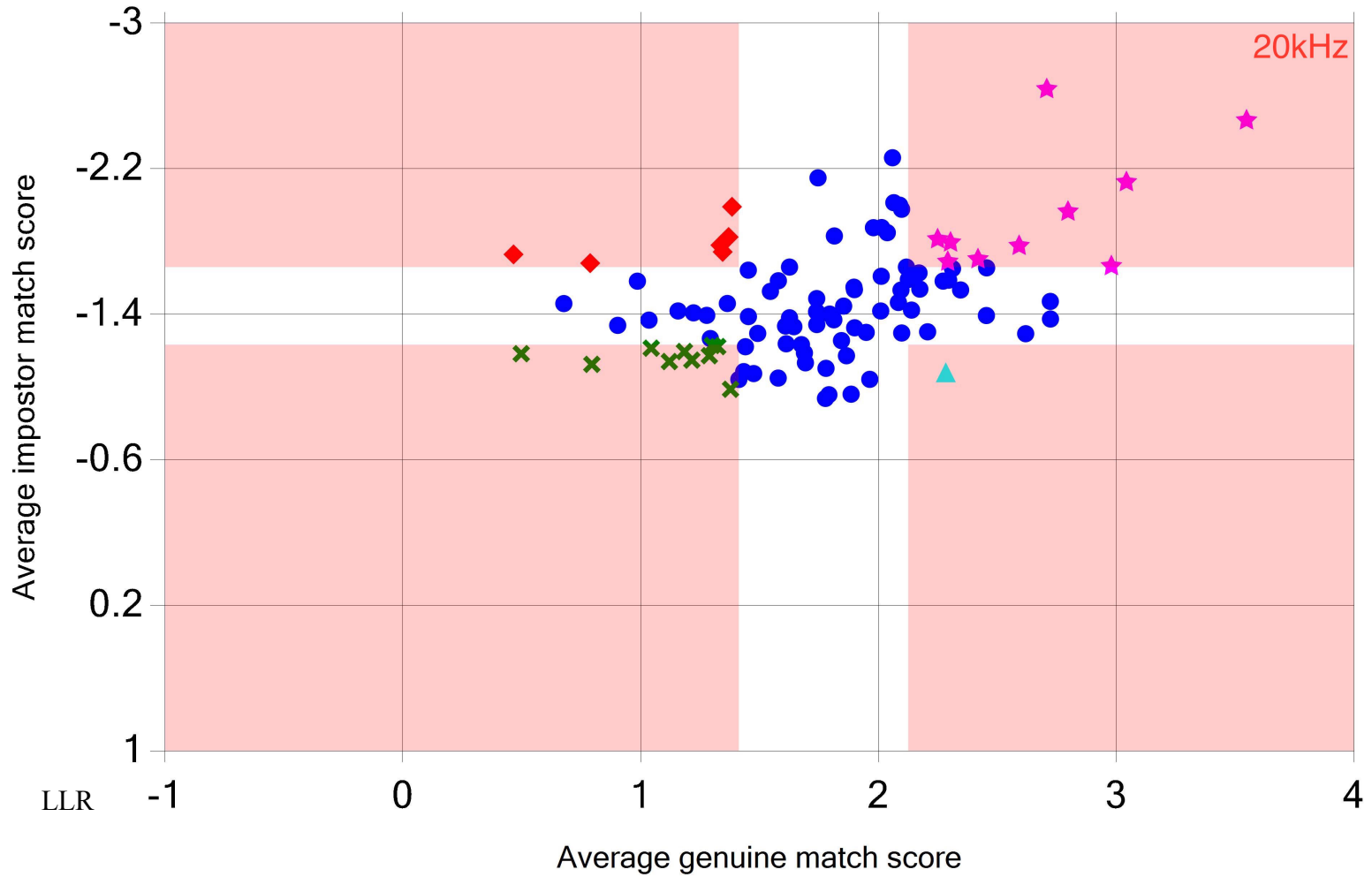


Sample Rate 24kHz, Frequency Bandwidth 0-12kHz



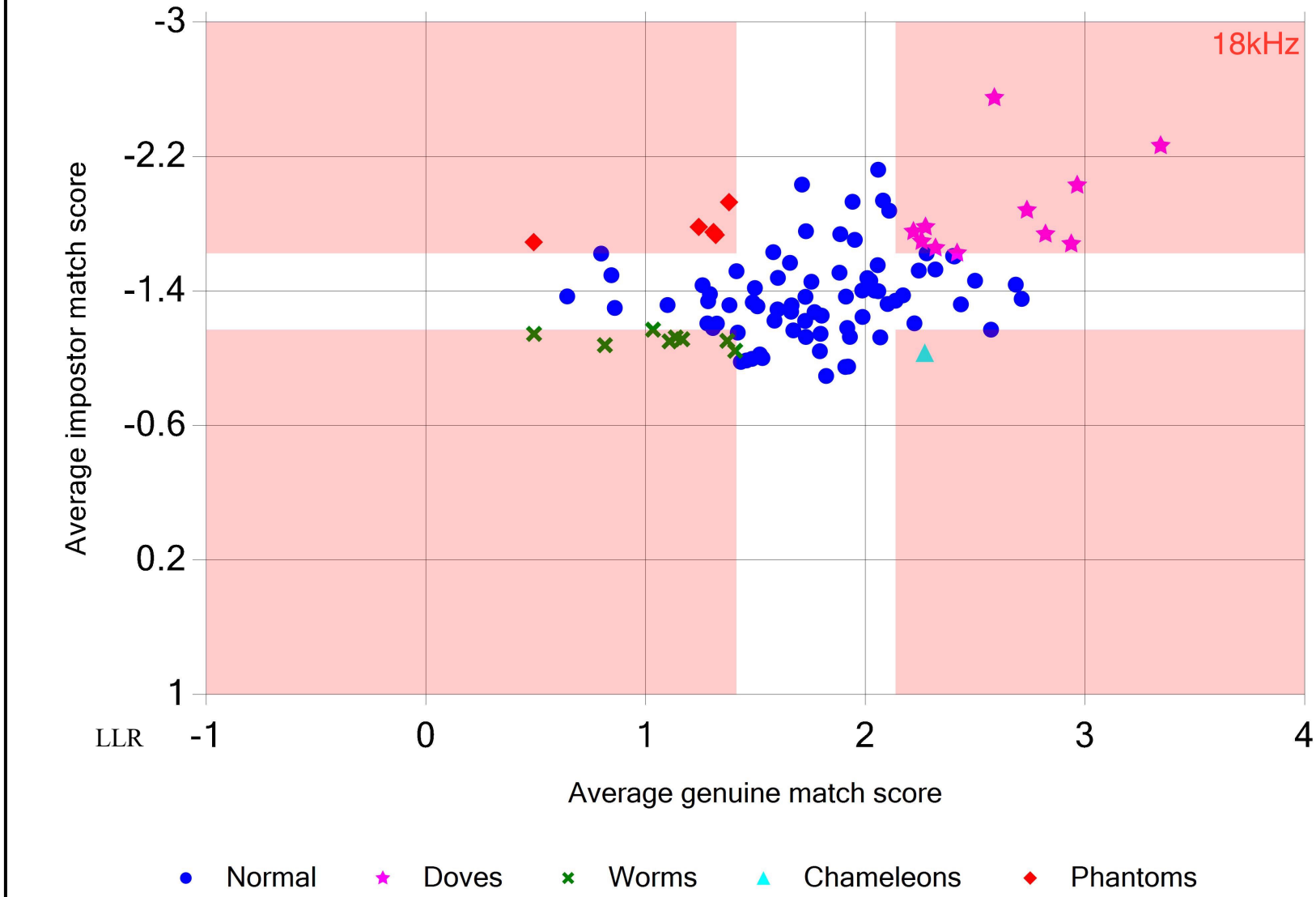


Sample Rate 20kHz, Frequency Bandwidth 0-10kHz

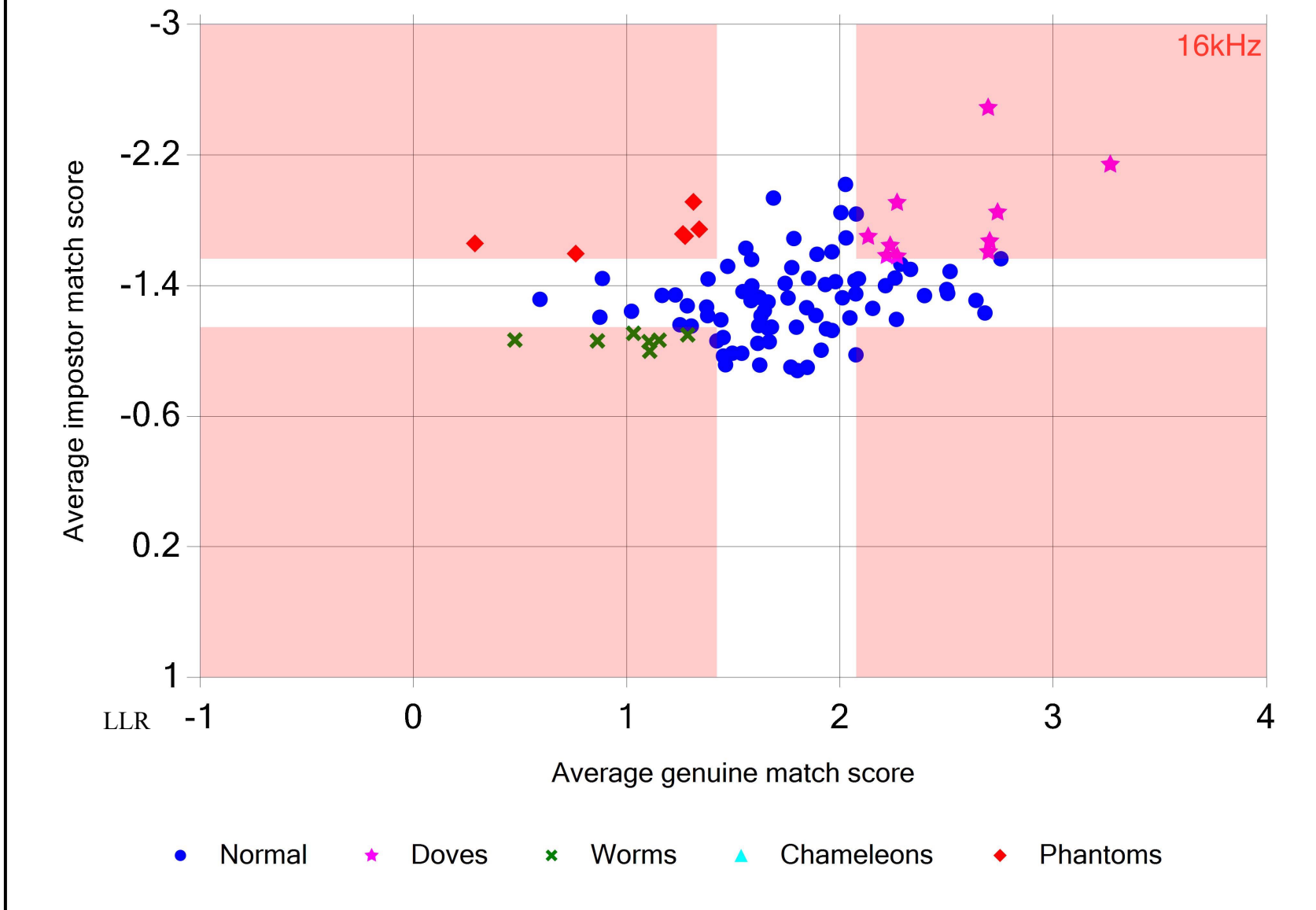


- Normal
- ★ Doves
- ✕ Worms
- ▲ Chameleons
- ◆ Phantoms

Sample Rate 18kHz, Frequency Bandwidth 0-09kHz

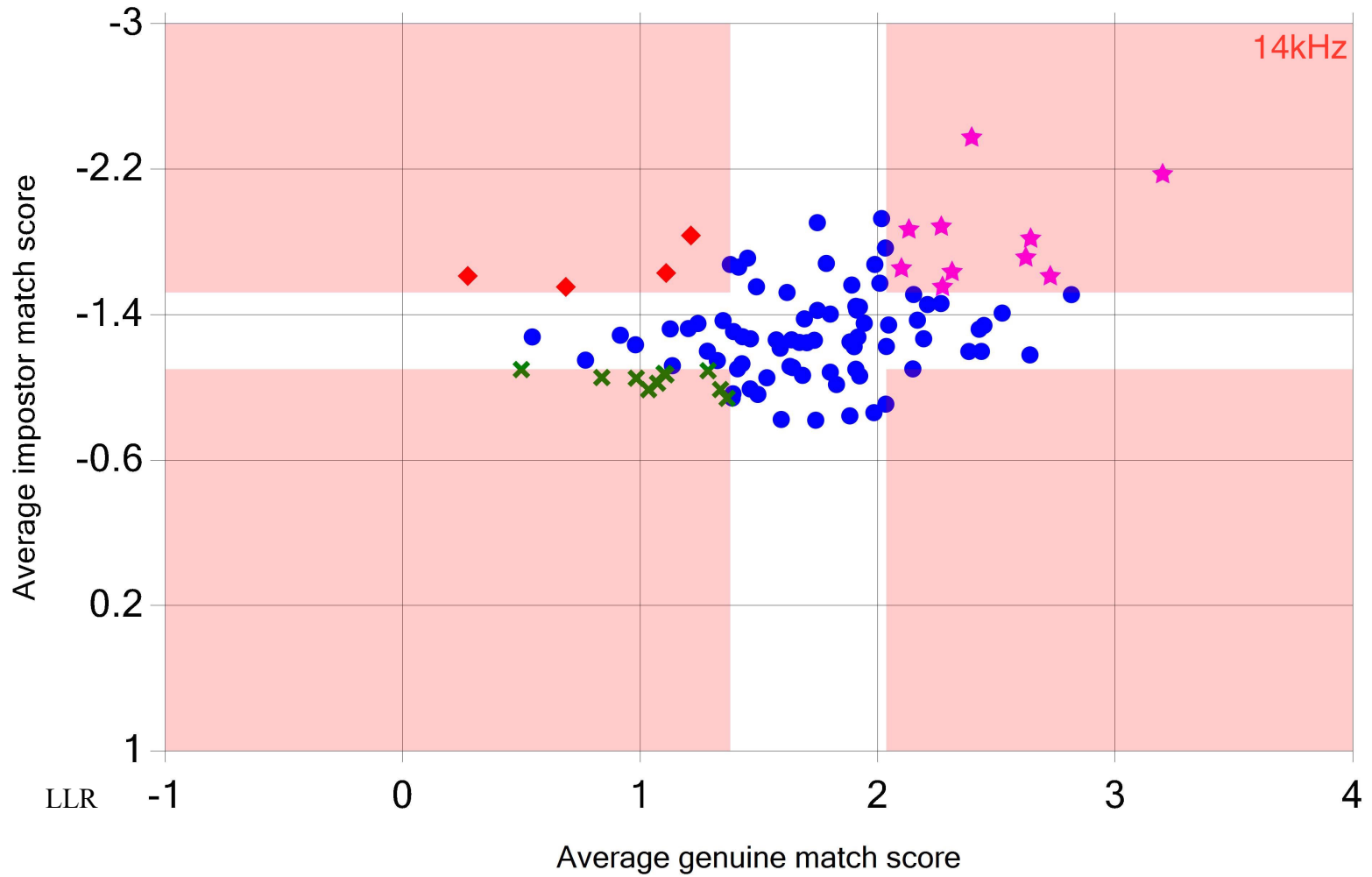


Sample Rate 16kHz, Frequency Bandwidth 0-08kHz



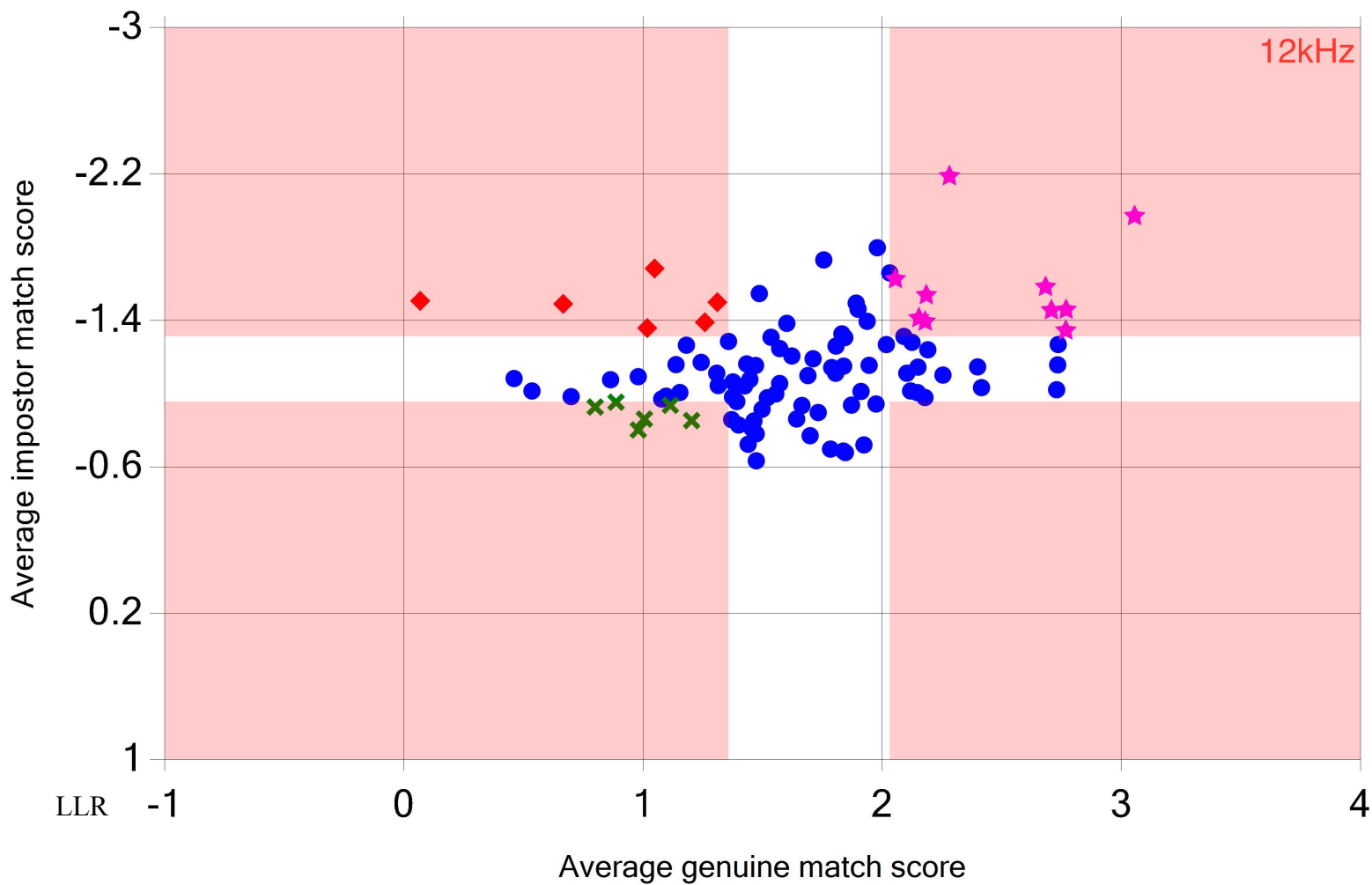


Sample Rate 14kHz, Frequency Bandwidth 0-07kHz



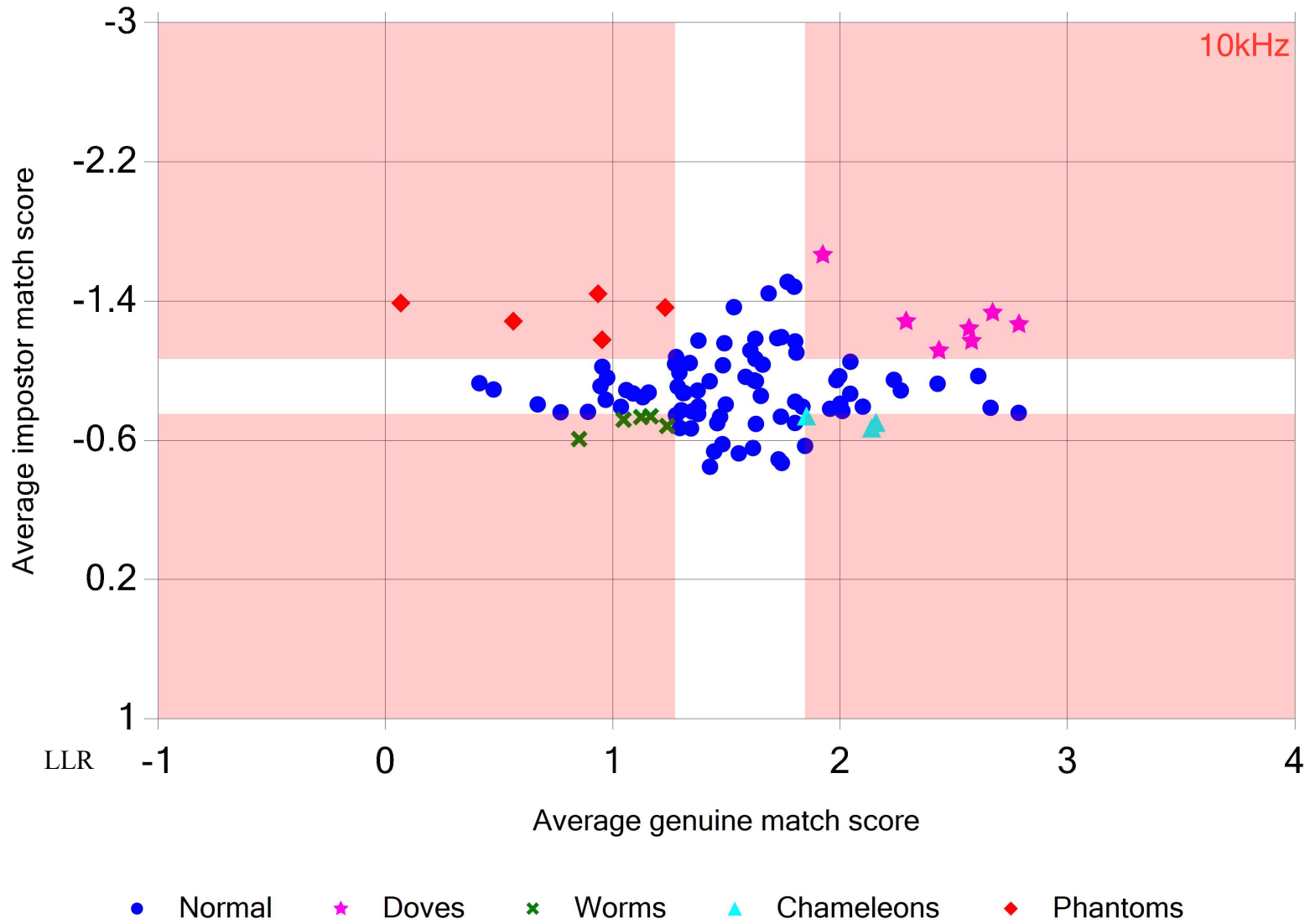
- Normal
- ★ Doves
- × Worms
- ▲ Chameleons
- ◆ Phantoms

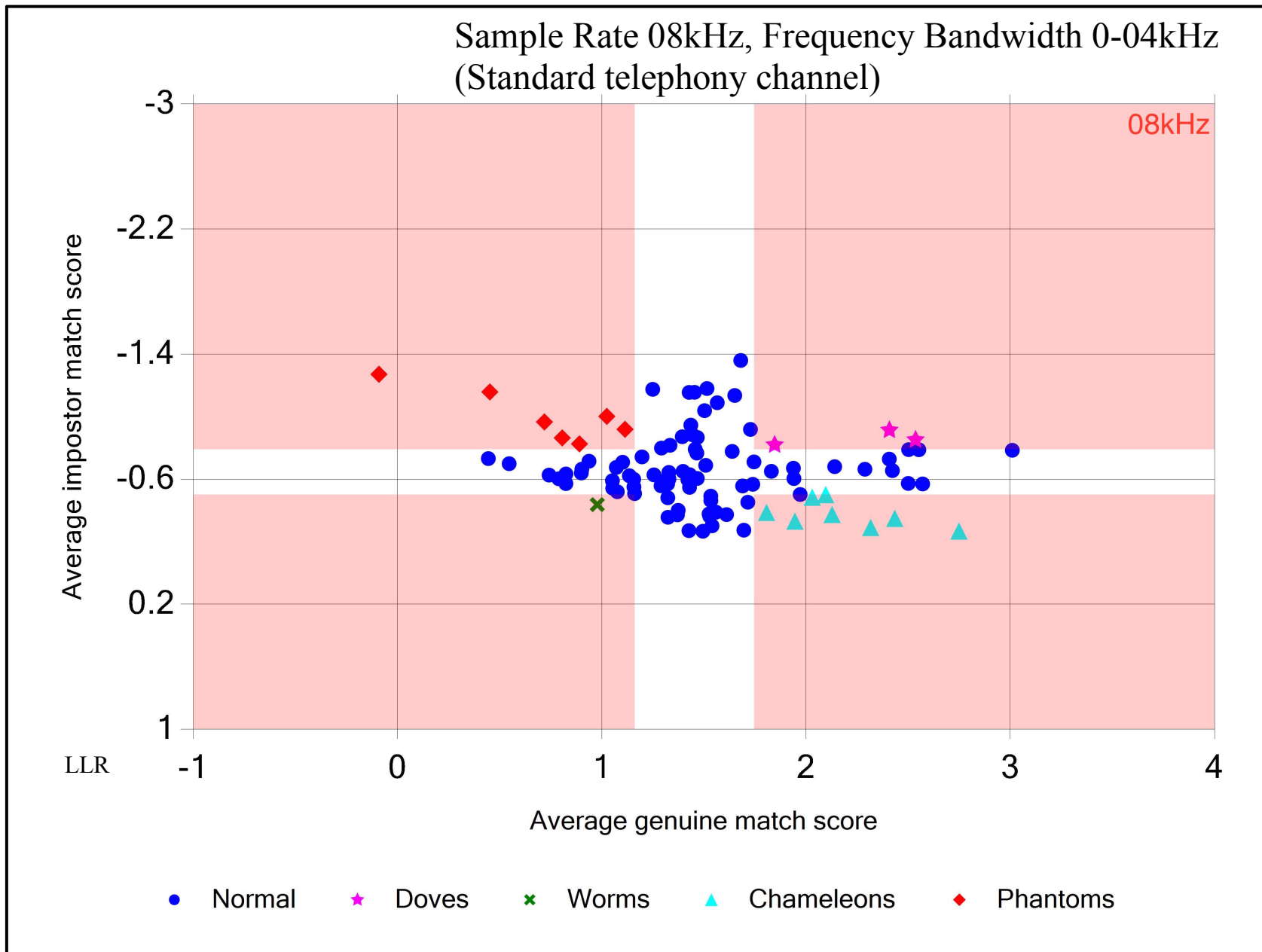
Sample Rate 12kHz, Frequency Bandwidth 0-06kHz



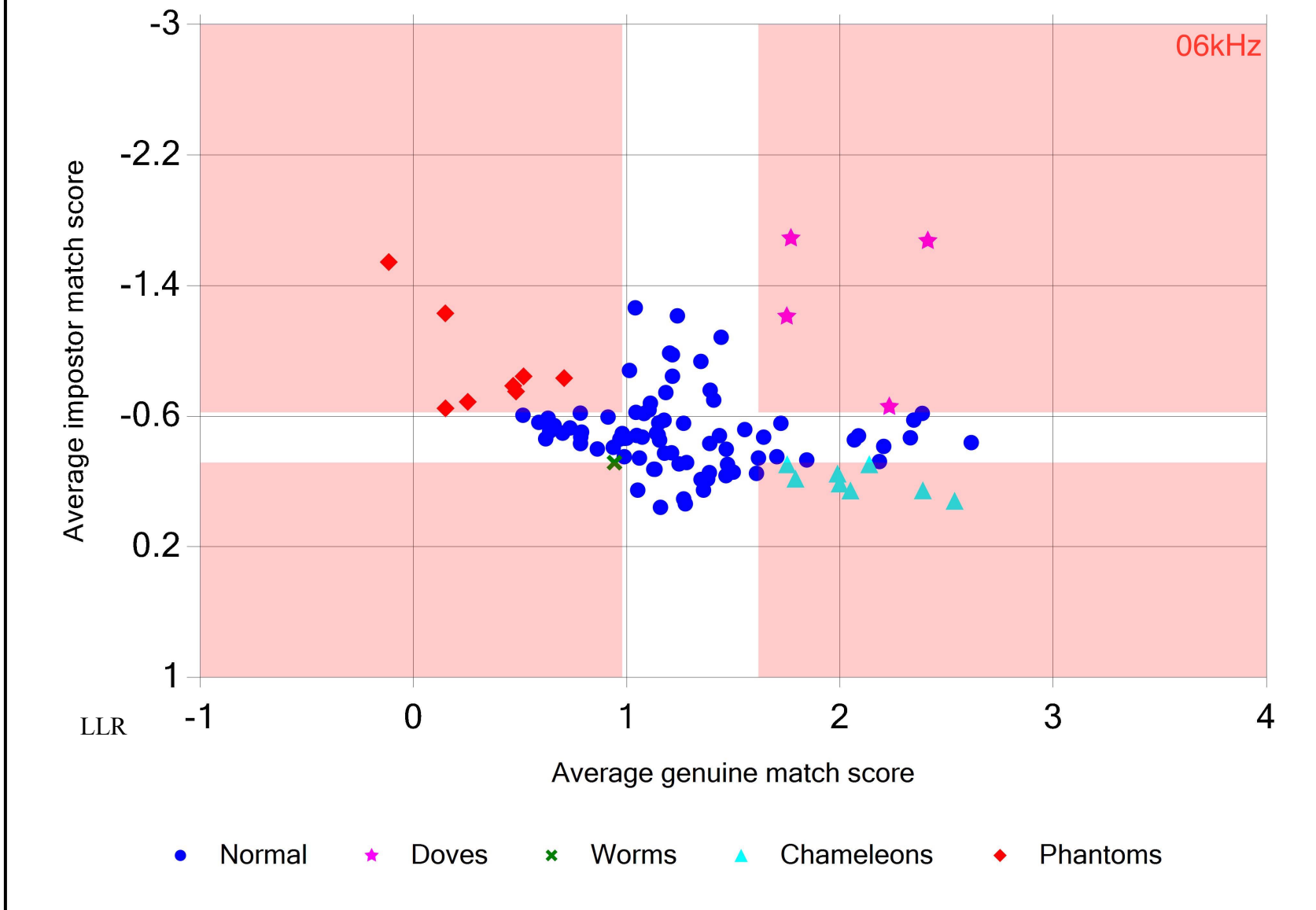
- Normal
- ★ Doves
- × Worms
- ▲ Chameleons
- ◆ Phantoms

Sample Rate 10kHz, Frequency Bandwidth 0-05kHz

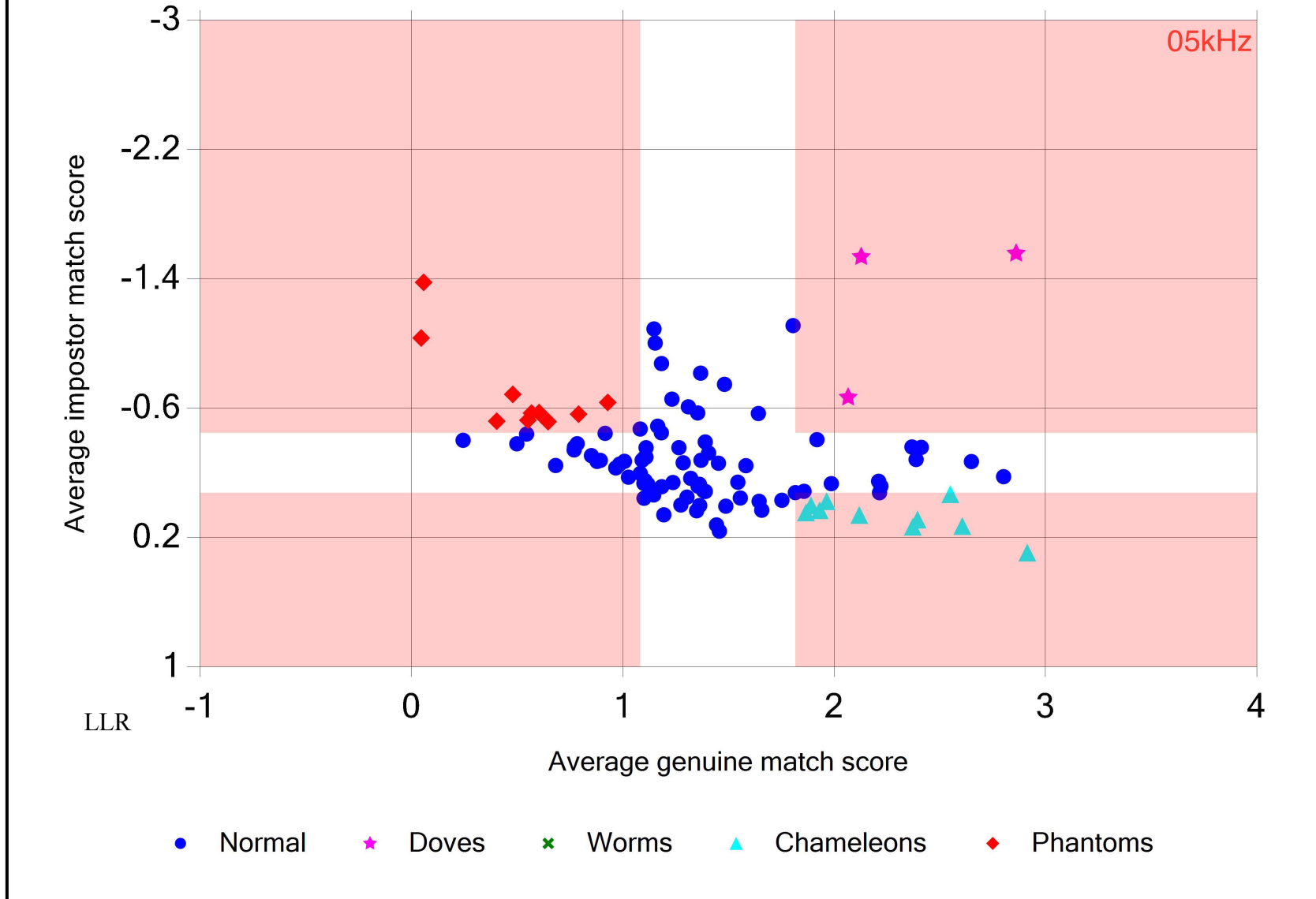




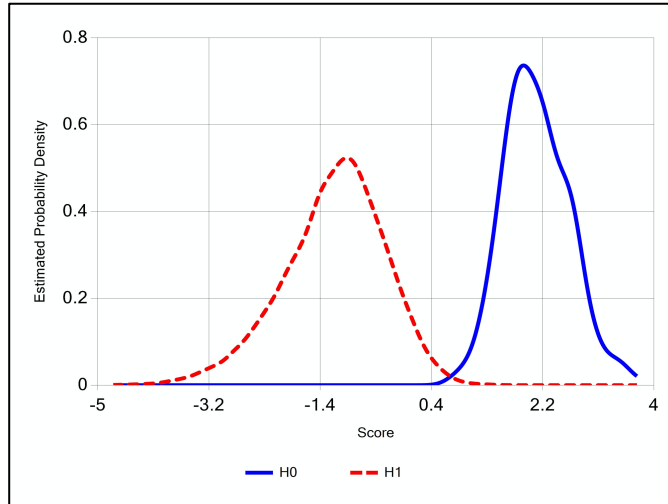
Sample Rate 06kHz, Frequency Bandwidth 0-03kHz



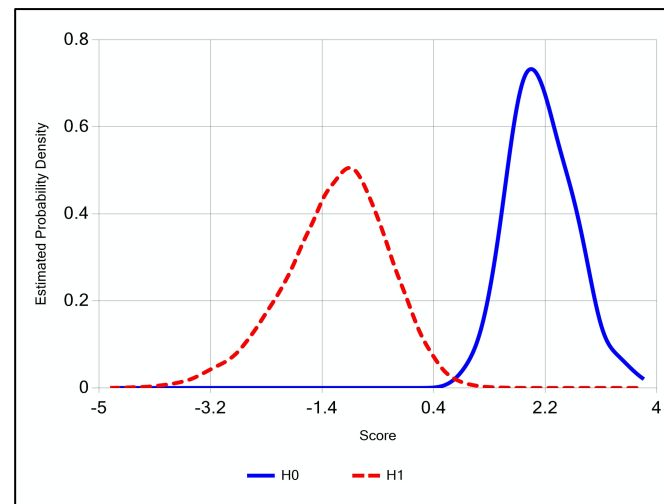
Sample Rate 05kHz, Frequency Bandwidth 0-2.5kHz



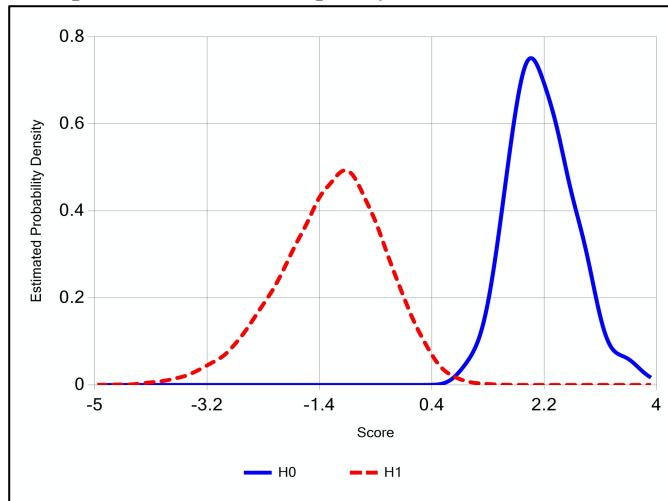
Sample Rate 32kHz, Frequency Bandwidth 0-16kHz



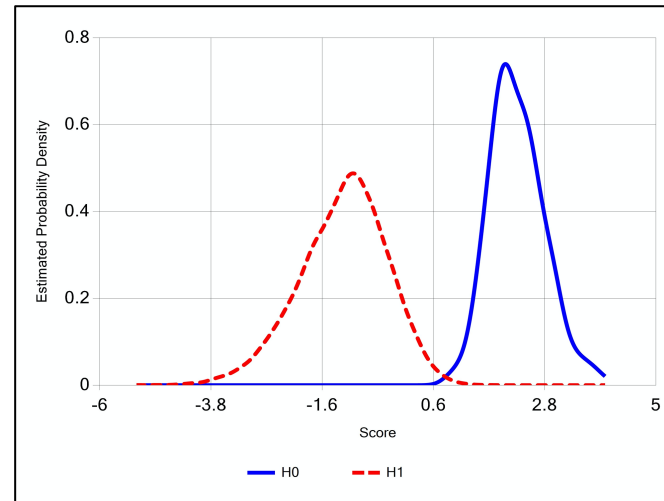
Sample Rate 30kHz, Frequency Bandwidth 0-15kHz



Sample Rate 28kHz, Frequency Bandwidth 0-14kHz

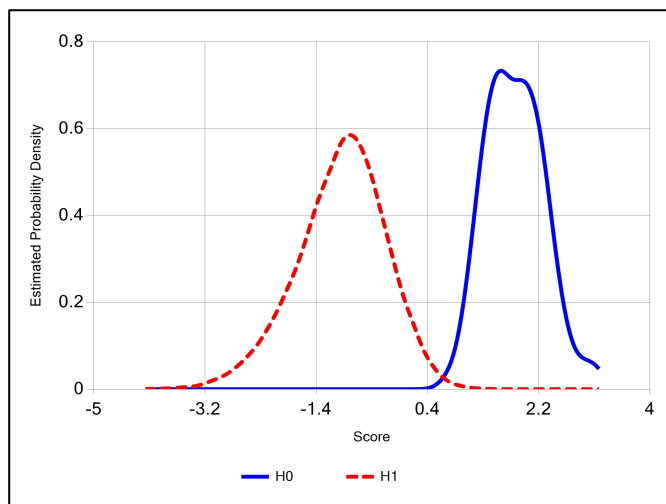


Sample Rate 26kHz, Frequency Bandwidth 0-13kHz

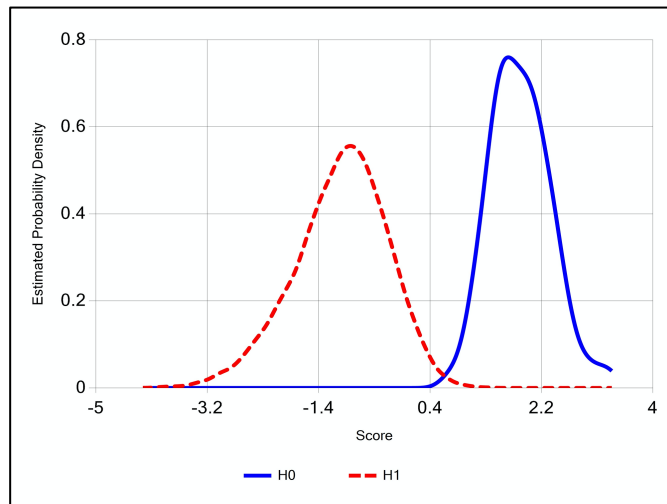


Negligible differences in LR performance observed at a system level, for higher frequency bandwidth settings. Individual speaker performance was affected (zoo plot position).

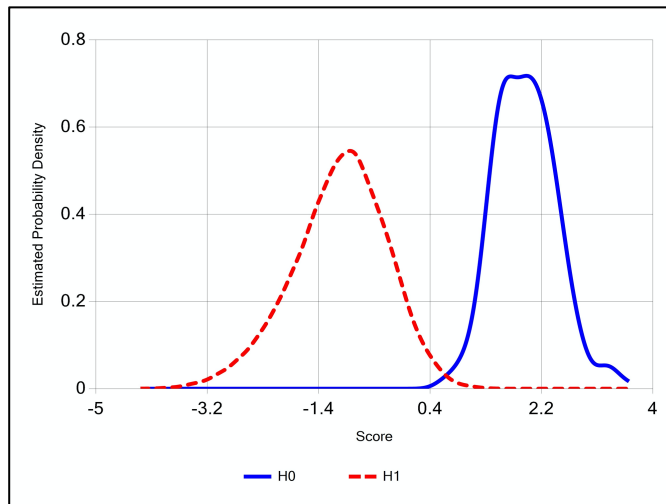
Sample Rate 24kHz, Frequency Bandwidth 0-12kHz



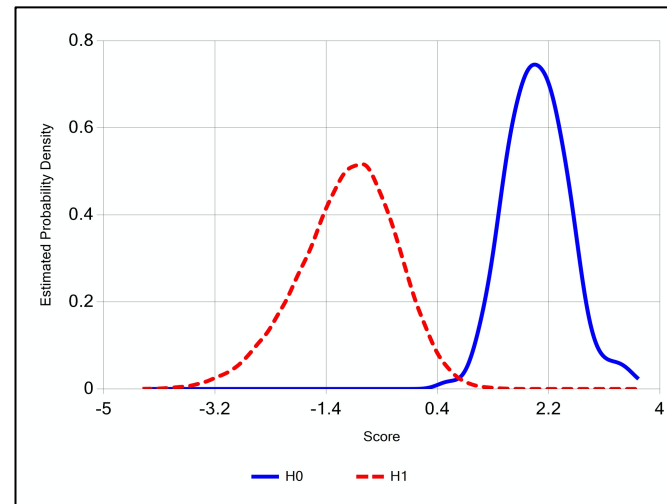
Sample Rate 22kHz, Frequency Bandwidth 0-11kHz



Sample Rate 20kHz, Frequency Bandwidth 0-10kHz

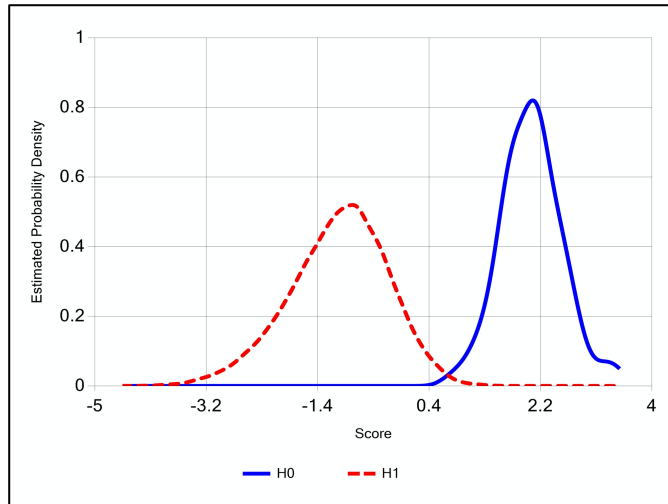


Sample Rate 18kHz, Frequency Bandwidth 0-09kHz

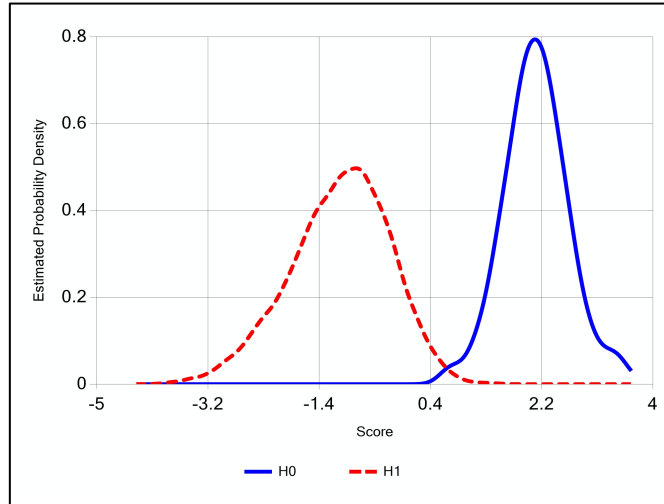




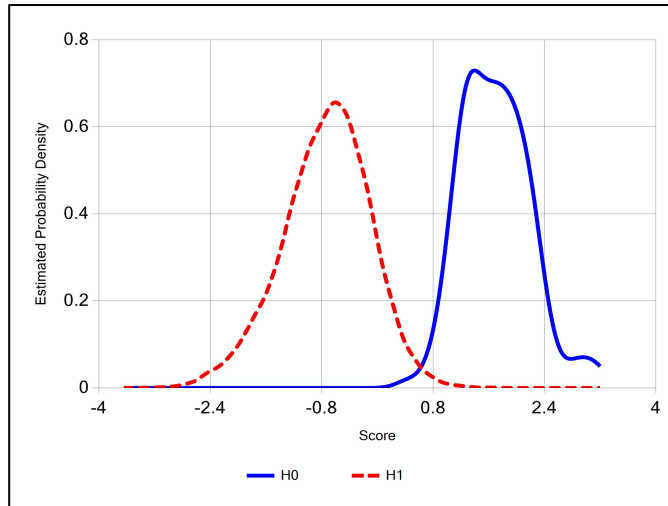
Sample Rate 16kHz, Frequency Bandwidth 0-08kHz



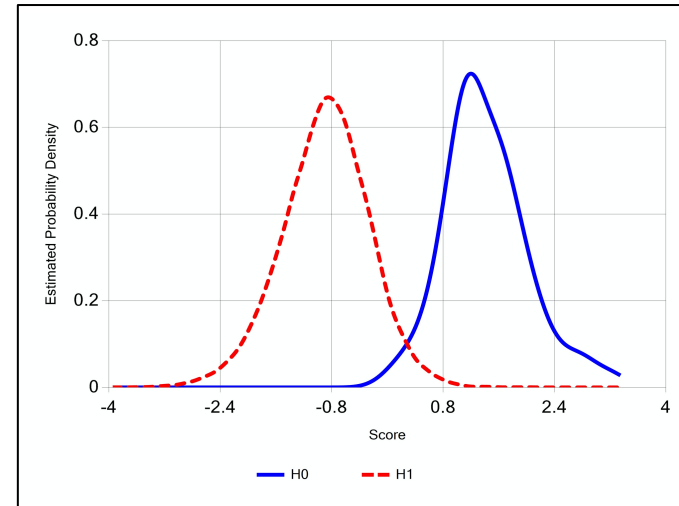
Sample Rate 14kHz, Frequency Bandwidth 0-7kHz



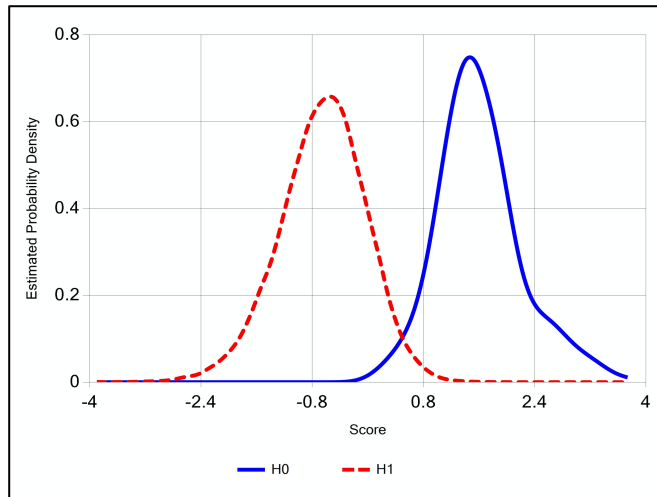
Sample Rate 12kHz, Frequency Bandwidth 0-6kHz



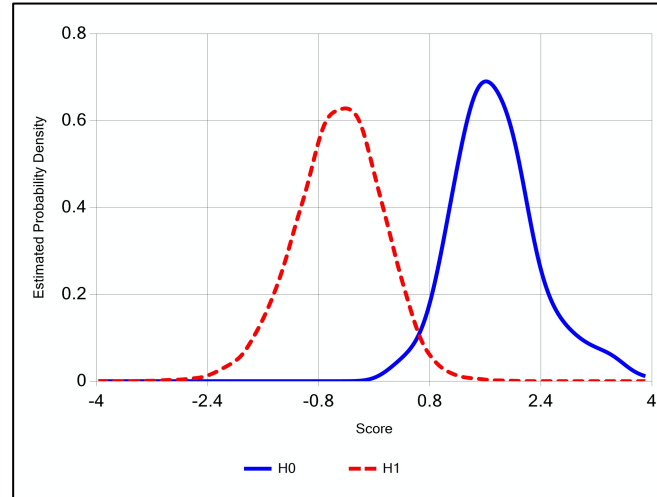
Sample Rate 10kHz, Frequency Bandwidth 0-5kHz



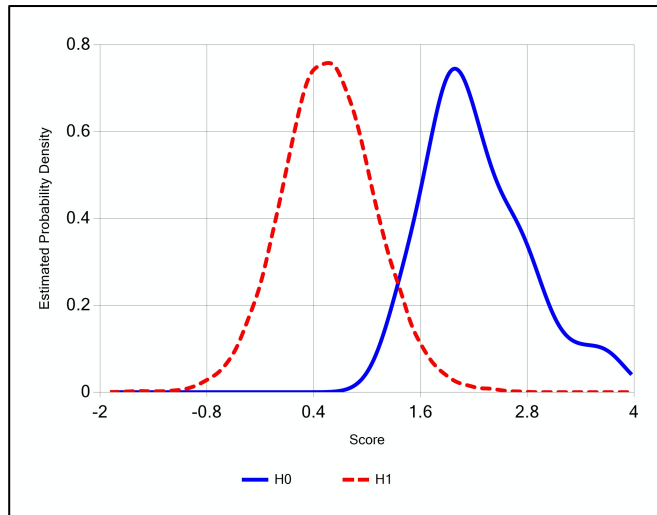
Sample Rate 08kHz, Frequency Bandwidth 0-04kHz



Sample Rate 07kHz, Frequency Bandwidth 0-3.5kHz



Sample Rate 06kHz, Frequency Bandwidth 0-3kHz



Significant shifts in LR performance observed at 0-7kHz and 0-6kHz (i.e. sub telephony channel)

# Appendix F

## Waves IR-L. Summary of reverberation settings

### Car Interiors

#### Ford Econoline 150

Capture Date: October 13, 2003  
Location: Nashville, Tennessee  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Locations: 1  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



#### Lincoln Navigator

Capture Date: October 13, 2003  
Location: Nashville, Tennessee  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Locations: 1  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



#### Tiled Bathroom

Capture Date: August 30, 2003  
Location: Tel Aviv, Israel  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



#### Livingroom

Capture Date: February 21, 2004  
Location: Tel Aviv, Israel  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Locations: 1  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



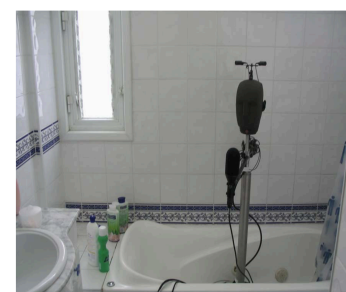
#### Kitchen

Capture Date: February 21, 2004  
Location: Tel Aviv, Israel  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Locations: 1  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



#### Bathroom

Capture Date: February 21, 2004  
Location: Tel Aviv, Israel  
Source Type: Genelec S30D  
Source Locations: Center  
Mic Locations: 1  
Mic Setup: Neumann SKM-140 (ORTF setup), Soundfield SPS-422B, Neumann KU-100 dummy head



# Appendix G

---

## **ASR Default system settings**

### **OWR Vocalise 1.5.0.1190\***

#### **GMM UBM System**

Default normative data unless otherwise stated.

MFCC settings applied 32 Gaussians, 13 Features, 24 filters, delta, CMS, symmetric, 10 train cycles.

\*For system consistency, tests completed using previous versions of Vocalise - e.g. Beta (2012), 1.3.0.607 (2013), 1.4.0.599 (2014), 1.4.0.651 (2014-15) and 1.5.0.1175 (2015) were either re run or validated on version 1.5.0.1190 (the last release of Vocalise version 1).

### **OWR iVocalise Version 2.5.0.1583 (2017B)\*\***

#### **I-Vector UBM, TV, LDA+PLDA System (default normative data unless otherwise stated)**

MFCC Settings: 13 features, 2 deltas, 24 filters, 1,024 gaussians

Total Variability: 10 train cycles, 400 dimensions

PLDA: 200 dimensions 10 train cycles

\*\*For system consistency, experiments completed on previous versions of iVocalise - e.g. 2.1.0.1366, 2.4.0.1547, or (2017a) were checked using version 2.5.0.1583.

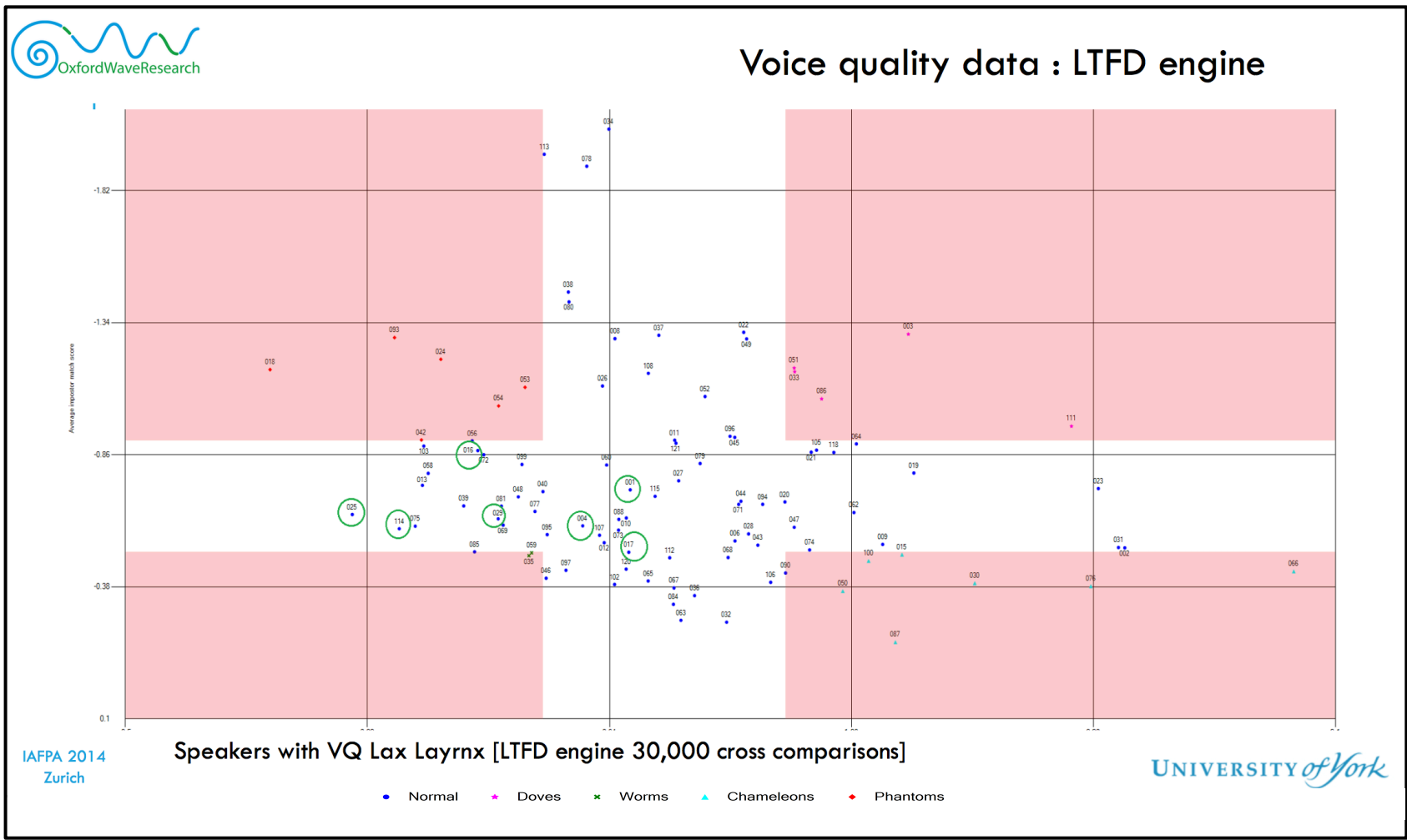
# Appendix H

---

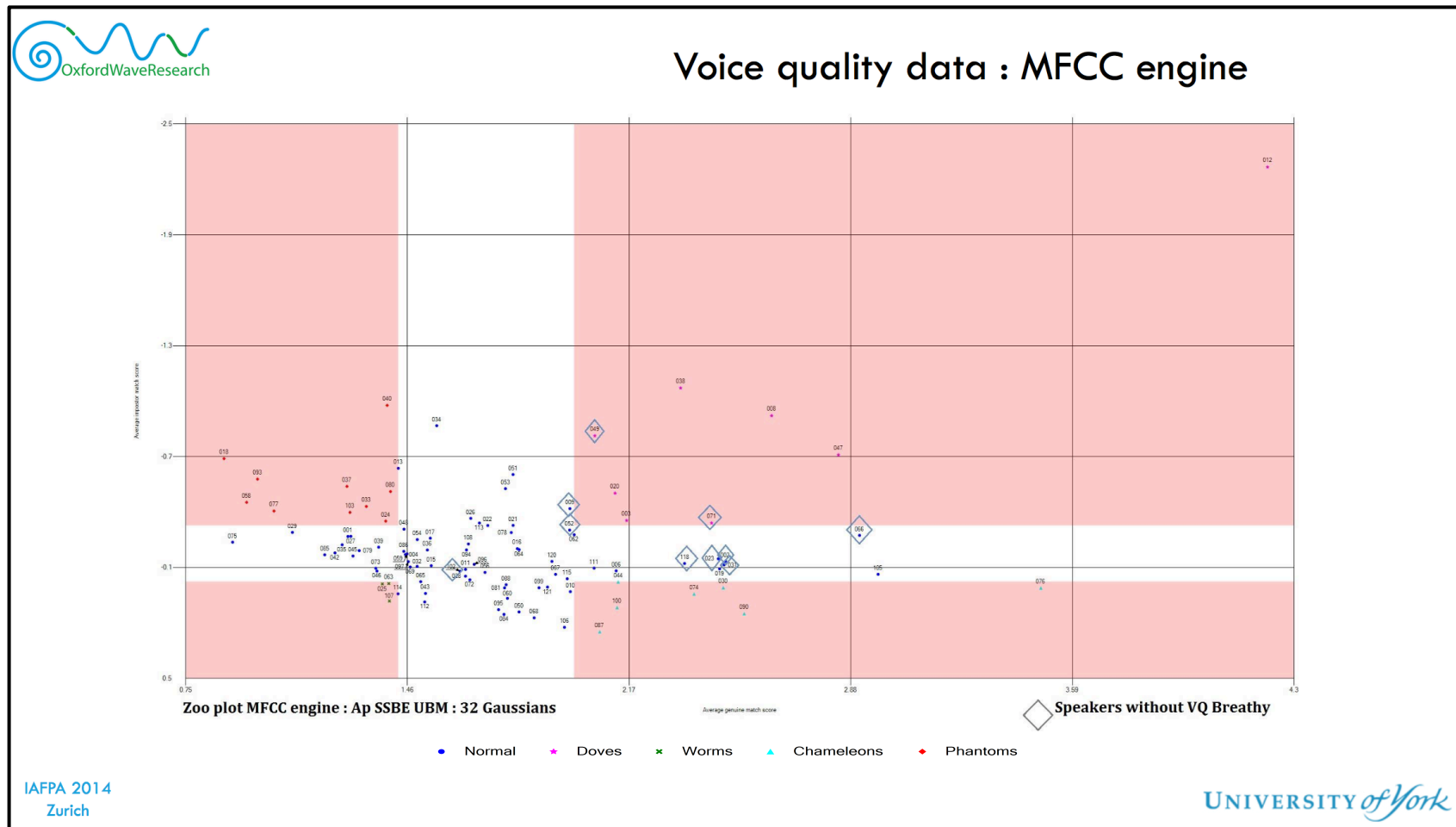
This section documents two slides that were presented at IAFPA (2014) by the author.

From: Alexander, A., Forth, O., Nash, J. and Yager, N. (2014). *Zoo plots for speaker recognition with tall and fat animals*.

Appendix H: Figure 1: OWR Vocalise LTFD. VQ analysis. Speakers with VQ lax larynx



Appendix H: Figure 2: OWR Vocalise MFCC. VQ analysis. Speakers without VQ breathy



# Appendix I

**Appendix I: Table 1:** Poddar, Sahidullah and Saha Tables (2015).  
Summary of net duration research to 2015 and key (below)

Results and comparison of i-vector based ASV techniques and advancements in short-duration condition.							
#Ref.	Modeling Methodology	Feature	Database (SRE)	Task	Length Train-Test	EER [%]	min DCF
[45]	JFA	MFCC	NIST '08	short2-short3	10s - 10s	21.17	0.0738
	TV+LDA+WCCN					21.56	0.0741
	TV+SDNAP+WCCN					20.84	0.0737
	TV+GPLDA					20.34	0.0762
[7]	Normalized i-vector	MFCC	NIST '10	core-core	10s - 10s	14.68	0.063
[80]	TV+GPLDA matched dev. data	MFCC	NIST '08	Short2-short3	10s - 10s	16.04	0.0679
	TV+HTPLDA matched dev. data					13.67	0.0639
[51]	TV+GPLDA (dev data: 10s segments)	MFCC	NIST '08	core	10s - 10s	10.70	0.0518
	TV+GPLDA (dev data: full segments)					11.77	0.0657
	TV+GPLDA (dev data: 10s+full segments)					9.79	0.0462
	TV+GPLDA (Fusion: 10s, full, 10s+full)					8.19	0.0442
[46]	TV+GPLDA (max. likelihood)	MFCC	NIST '10	8conv-10sec	Long-10s	9.89	-
	TV+GPLDA (minimax)					7.99	-
[90]	TV+GPLDA	MFCC	NIST '08	short2-short3	10s - 10s	15.07	0.0673
	TV+WCCN+GPLDA					14.99	0.0674
	TV+WCCN+LDA+GPLDA					15.80	0.0664
	TV+WCCN+SNLDA+GPLDA					15.40	0.0661
	TV+SUVN[LDA]+GPLDA					14.75	0.0618
TV+SUVN[SNLDA]+GPLDA	14.73	0.0620					
[92]	TV+GPLDA	MFCC	NIST '10	core-core	Long - 10s	10.4	0.0438
	TV+GPLDA (Modified Prior)					9.9	0.0464
[40]	TV+GPLDA	MFCC	NIST '08	short2-short3	10s - 10s	13.47	0.0635
	GMM-UBM					16.62	0.0700
[94]	TV+LDA+WCCN	MFCC	NIST '03	evaluation plan	Long - 10s	5.81	0.1090
		MPDSS+MFCC				5.56	0.1048
		RMFCC+MFCC				5.78	0.1087
		DCTILPR+MFCC				5.33	0.0971
[95]	TV+LDA+WCCN	MFCC	NIST '03	evaluation plan	Long - 10s	5.81	0.1090
	AFA+LDA+WCCN					4.92	0.0882
	Fusion: AFA, TV					4.29	0.0802



Poddar, Sahidullah and Saha (2015) Key.

Ref 45: Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: 'I-vector based speaker recognition on short utterances', Interspeech, 2011, pp.2341–2344
Ref 7: Mandasari, M.I., McLaren, M., van Leeuwen, D.A.: 'Evaluation of i-vector speaker recognition systems for forensic application.', Interspeech, 2011, pp.21–24
Ref 80: Kanagasundaram, A., Vogt, R.J., Dean, D.B., Sridharan, S.: 'PLDA based speaker recognition on short utterances', Proc Odyssey, 2012
Ref 51: Sarkar, A.K., Matrouf, D., Bousquet, P.M., Bonastre, J.F.: 'Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification.', Interspeech, 2012
Ref 46: Hautamäki, V., Cheng, Y.C., Rajan, P., Lee, C.H.: 'Minimax i-vector extractor for short duration speaker verification.', Interspeech, 2013, pp.3708–3712
Ref 90: Kanagasundaram, A., Dean, D., Sridharan, S.: 'Improving PLDA speaker verification with limited development data', Proc ICASSP, 2014, pp.1665–1669
Ref 92: Hong, Q., Li, L., Li, M., Huang, L., Wan, L., Zhang, J.: 'Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system.', Interspeech, 2015, pp.1037–1041
Ref 40: Poddar, A., Sahidullah, M., Saha, G.: 'Performance comparison of speaker recognition systems in presence of duration variability.', Proc IEEE INDICON, 2015, pp.1–6
Ref 94: Das, R.K., Jelil, S., Prasanna, S.M.: 'Significance of constraining text in limited data text-independent speaker verification', Proc SPCOM, 2016, pp.1–5
Ref 95: Mamodiya, S., Kumar, L., Das, R.K., Prasanna, S.M.: 'Exploring acoustic factor analysis for limited test data speaker verification', Proc TENCON, 2016, pp.1397–1401

# Appendix J

---

Additional experiment re Chapter 11. Test tone tables x3.

**Appendix J: Table 1:** Mean test tone frequencies (Praat, baseline)

Tone 1	Tone 2	Tone 3	Tone 4
546.60059	1202.388516	2598.977815	3318.984447
546.797115	1202.451478	2599.016631	3325.704933
546.878872	1202.477724	2599.033655	3328.553715
546.733221	1202.431067	2599.004172	3323.500019
546.520082	1202.362632	2598.96199	3316.288651
546.297263	1202.291804	2598.917013	3308.900544
546.129906	1202.23773	2598.882992	3303.479487
546.066561	1202.217015	2598.869952	3301.422002
546.094571	1202.226449	2598.876853	3302.331065
546.313633	1202.296846	2598.920203	3309.445764
546.807269	1202.454838	2599.019287	3326.036962
547.466224	1202.663153	2599.151369	3349.779744
548.368012	1202.943726	2599.331311	3384.694425
549.455919	1203.27593	2599.547196	3400.898613
550.989765	1203.728033	2599.84265	3400.410311
552.720663	1204.211736	2600.165063	3400.495909

**Appendix J: Table 2:** Mean test tone frequencies (Praat, .mp3 CBR 8kbps)

Tone1	Tone 2	Tone 3	Tone 4
540.705545	1200.714588	2580.75331	Removed
540.320115	1200.572588	2579.979998	
540.19131	1200.41268	2580.000793	
540.353264	1200.534735	2580.146547	
540.273517	1200.645247	2580.380521	
540.465987	1200.526304	2580.27643	
540.545455	1200.395059	2580.382642	
540.271766	1200.572892	2580.353637	
540.645923	1200.751397	2580.333727	
541.518586	1201.024833	2580.752848	
540.165862	1200.730631	2580.200643	
540.79301	1200.442974	2580.837912	
540.438248	1200.674219	2580.33718	
540.881064	1200.635909	2580.246591	
540.24833	1200.742071	2580.433598	
541.295054	1200.902367	2580.468051	

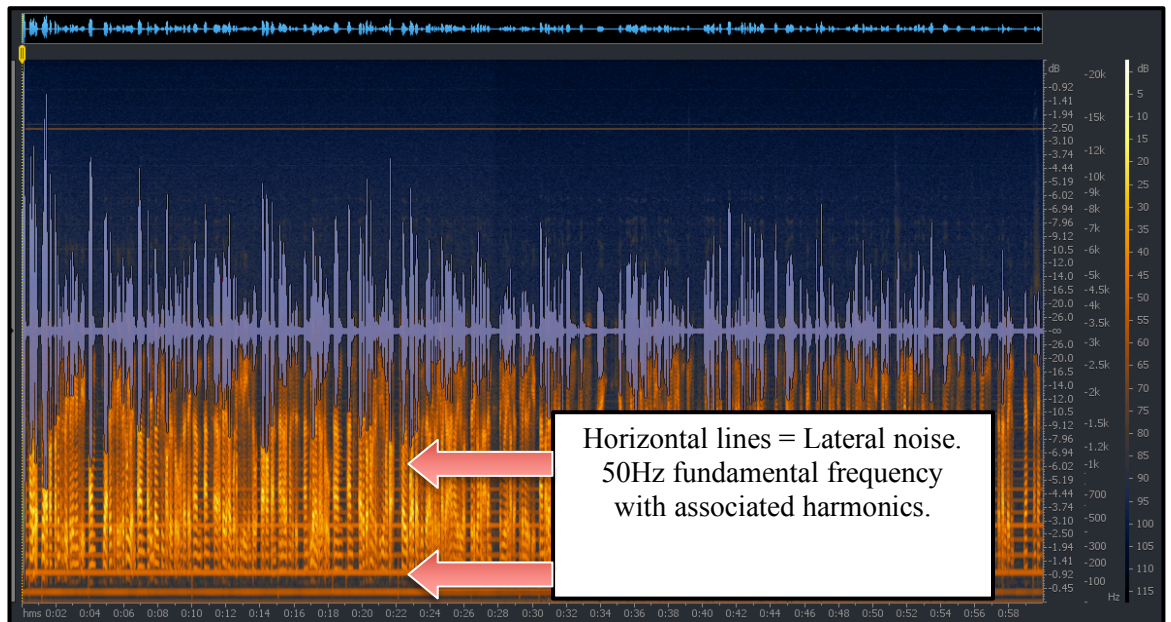
**Appendix J: Table 3:** Mean test tone frequencies (Praat, Speex ‘8’)

Test Tone 1	Test Tone 2	Test Tone 3	Test Tone 4
563.538229	1207.048712	2599.832934	3400.124593
564.292315	1206.44798	2602.39524	3400.663633
563.525377	1207.23385	2598.743349	3400.563523
564.606012	1207.527922	2600.9558	3400.948435
565.771174	1208.138565	2601.056791	3400.061066
563.88503	1206.852521	2601.21012	3401.504476
565.105173	1206.458962	2601.564089	3400.846961
563.311372	1207.891123	2601.775288	3401.158583
565.492157	1207.56348	2601.065877	3400.417626
564.97023	1207.401238	2601.252136	3400.493404
562.829164	1207.44742	2600.574466	3400.612519
563.074696	1206.327041	2601.436151	3399.241884
563.342912	1205.834619	2600.941448	3399.430941
561.05416	1207.680018	2600.563622	3400.751288
561.901265	1205.745303	2602.069995	3400.263697
562.036373	1205.13095	2599.480283	3400.498226

# Appendix K

Spectrogram of speaker 012, DyViS – demonstrating noise (mains hum).

**Appendix K: Figure 1:** iZotope RX Spectrogram. DyViS Speaker 012, task 2



# Appendix L

---

Parallels (Support Ticket) #2713244 18:47 23/2/20

John,

Thank you for contacting Parallels Support. This email is in reference to the query about audio files in Parallels Desktop.

We would like to inform you that, Parallels Desktop provides sound output for the guest operating system by emulating a virtual sound device inside your virtual machine. On the host operating system side, Parallels Desktop uses sound in a similar manner to any other application it will not affect or corrupt the audio files stored under a guest operating system.

Parallels Desktop uses a special type of the virtualization: a hardware-assisted full hardware virtualization that relies on the Intel VT-X technology and allows simulating the whole computer with both its hardware and software. Please refer to the 'blog below for further information about audio settings in Parallels Desktop\*.

<https://www.parallels.com/blogs/parallels-sound-troubleshooting/>.

If you require any further assistance, please reply to this email.

Thank you

Angamuthu Mahadevan

Parallels Technical Support

\*Please note that this 'blog refers predominantly to sound playback between the virtual and host OS.

# Abbreviations

---

AES	Audio Engineering Society
AMR	Adaptive Multi-Rate
ASR	Automatic/Automated Speaker Recognition
ASV	Automatic/Automated Speaker Verification
BPF	Band Pass Filter
CLEAVER	Cluster Estimation And Versatile Extraction of Regions (OWR). Speaker segmentation/diarisation application.
Clr	Cost of Log Likelihood Ratio
CODEC	CODing and/or DECoding algorithm, usually for audio or video
COTS	Commercial Off The Shelf
CSV	Comma Separated Values
CTEST	‘Contest’ database of recordings (see also SPOKE)
DET	Detection Error Trade off
DL	Deep Learning
DNN	Deep Neural Networks
DOCC	Damped Oscillator Cepstral Coefficients
DSP	Digital Signal Processing
DyViS	Dynamic Variability in Speech (Cambridge speech corpus)
EEG	Equal Error Graph
EER	Equal Error Rate
F0, F1, F2, F3, F4...	Formants, numbered from fundamental frequency (F0) upwards
FAR	False Accept Rate
FFT	Fast Fourier Transform
FNR	False Negative Rate
FPR	False Positive Rate
FRR	False Reject Rate
FSC	Forensic Speaker Comparison
FSS	Forensic Speech Science
GMM	Gaussian Mixture Model
GSM	Global System for Mobile communications
GT	Ground Truth
GUI	Graphic User Interface
HASR	Human Assisted Speaker Recognition
HPF	High Pass Filter
IAFPA	International Association of Forensic Phonetics and Acoustics

IR	Impulse Response
ISCA	International Speech and Communication Association
I VOCALISE	Voice Comparison & Analysis of the Likelihood of Speech Evidence (i-vector version).
LDA	Linear Discriminant Analysis
LEA	Law Enforcement Agency
LPF	Low Pass Filter
(L)LR	(Log) Likelihood Ratio
LTFD	Long Term Formant Distribution
MCD	Mel Cepstral Dynamics
MFCC	Mel Frequency Cepstral Coefficients
MIRS	Modified Impulse Response System
MSR	Microsoft Speaker Recognition (released 2013)
NB	Narrow Band telephony data 0-4kHz (8kHz SR)
NIST	National Institute of Standards and Technology (U.S.)
OS	Open Source
OWR	Oxford Wave Research
PDF	Probability Density Function (also LR Plot)
PLDA	Probabilistic Linear Discriminant Analysis
PTTR	Push To Talk Radio (‘walkie-talkies’ commonly used by LEAs/security/military)
ROC	Receiver Operating Characteristic
RT60	Reverb Time 60db (time for a reverberant sound to drop by 60db)
SAD	Speech Activity Detection (see also SD and VAD)
SD	Standard Deviation
SD	Speech Detection (see also VAD and SAD)
SM	Speaker Model (see also ‘Voice Print’)
SNR	Signal to Noise Ratio
SPARSE	Selective Processing of Annotated Regions of Speech Efficiently
SPEEX	An open source codec
SPOKE	Speech Obtained in Key Environments. Speech corpora (HO/OWR GUI & Database)
SR	Sample Rate
SRC	Sample Rate Conversion
SRE	Speaker Recognition Evaluation (NIST data set)
SSBE	Standard Southern British English
SWB	Super Wide Band, type of telephony data (50Hz to 14kHz)
T60	See RT60

TA	Test Audio
TN	True Negative
TP	True Positive
TV(M)	Total Variability (Matrix)
UBM	Universal Background Model (normative data)
VAD	Voice Activity Detection (see also SD and SAD)
VOCALISE	Voice Comparison & Analysis of the Likelihood of Speech Evidence. Version 1, GMM-UBM. See also iVocalise.
VOT	Voice Onset Time
VP	Voice Print (see also Speaker Model)
VPA(S)	Vocal Profile Analysis (Scheme)
VQ	Voice Quality
VLR	Vowel Like Region(s)
VTL	Vocal Tract Length
WAV	<u>Waveform</u> , audio file format
(S)WB	(Super) Wide Band (greater than narrow band, 0-4kHz, 8kHz SR).
XML	Extensible Mark-up Language



# Bibliography

---

- Abercrombie, D. (1967). *Elements of general phonetics* (reprinted 1982). Edinburgh University press.
- Ahmed, N. (1991). How I came up with the discrete cosine transform (in 1972). *Digital Signal Processing* (1), pp.4-5.
- Aitken, C.G.G. (1983). Statistics and forensic Science - a fruitful partnership. *Journal of the Forensic Science Society* 23(1), pp.3-4.
- Aitken, C.G.G. (1987). Statistics in forensic science. *Journal of the Forensic Science Society*, 27(2), pp.113-115.
- Aitken, C.G.G. and Stoney, D. (1991). *The use of statistics in forensic science*. Chapman and Hall. ProQuest E-book Central.
- Akula, A. and DeLeon, P. (2008). Compensation for room reverberation in speaker identification. Proceedings of the 16<sup>th</sup> European Signal Processing Conference. Lausanne, Switzerland.
- Akula, A., Apsingekar, V.R. and De Leon, P.L. (2009). Speaker identification in room reverberation using GMM-UBM. 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, pp.37-41.
- Al-Karawi, K., Al-Noori, A., Li, F. and Ritchings. T. (2015). Automatic speaker recognition system in adverse conditions. *Implication of noise and reverberation on system performance*. International Journal of Information and Electronics Engineering, 5(6), pp.423-427.
- Al-Noori, A. and Duncan, P. (2019). Robust speaker recognition in noisy conditions by means of online training with noise profiles. The Audio Engineering Society Journal 67(4). Cited from: [http://www.aes.org/journal/online/JAES\\_V67/4/](http://www.aes.org/journal/online/JAES_V67/4/).
- Alexander, A. (2005). *Forensic automatic speaker recognition using Bayesian interpretation for mismatched conditions*. Lausanne university, Switzerland. PhD thesis.
- Alexander, A. (2007). Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions. *International Journal of Speech Language and the Law* 14(1), pp.145-156.
- Alexander, A. and Forth, O. (2012). Blind speaker clustering using phonetic and spectral features in simulated and realistic police interviews. Presented at IAFPA, Santander, Spain, 2012.

- Alexander, A., Forth, O. and Tunstall, D. (2012). Music and noise fingerprinting and reference cancelling applied to forensic audio enhancement. AES 46<sup>th</sup> Conference, Denver, USA.
- Alexander, A., Forth, O. and Jessen, M. (2013). Speaker Recognition with Phonetic and Automatic Features using VOCALISE software. Presented at IAFPA, Tampa, Florida, 2013.
- Alexander, A., Forth, O., Nash, J. and Yager, N. (2014). Zoo plots for speaker recognition with tall and fat animals. OWR and York University, AICBT Ltd. Presented at IAFPA, Zurich, Switzerland, 2014.
- Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2016). VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), Bilbao, Spain.
- Alexander, A., Forth, O. and Kelly, F. (2018). Bio-Metrics software. 2018a. Version 1.8.0.710, manual.
- Atal, B.S. and Hanauer S.L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2), pp.637-655.
- Athulya, M., Vinayshankar and Sathidevi, P. (2017). Mitigating effects of noise in forensic speaker recognition. Department of ECE, National Institute of Technology, Calicut, India. IEEE, WiSPNET, 2017.
- Athulya, M. and Sathidevi, P. (2018). Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers. Electronics and Communication Engineering Department, National institute of technology, Calicut, India.
- Avila, A.R., Paja, M.O.S., Francisco, F.J, O'Shaughnessy, D. and Falk, T.H. (2015). Improving the performance of far-field speaker verification using multi-condition training: the case of GMM-UBM and i-vector systems. Proceedings of the Interspeech Conference, pp.1096-1100.
- Becker, T. (2007). The influence of intra-speaker variability in automatic speaker Identification. Presented at IAFPA, Plymouth, United Kingdom, 2007.
- Becker, T., Jessen M. and Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian Mixture Models. Proceedings of the Interspeech Conference 2008, Brisbane, Queensland, Australia.
- Becker, T., Jessen, M., Alsbach, S., Broß, F. and Meir, T. (2010). Automatic forensic comparison using recording adapted background models. Proceedings of the Audio Engineering Society

(AES) International Conference, Hillerød, Denmark.

Becker, T., Jessen, M., Broß, F. and Meir, T. (2011). The effect of MP3 compression on automatic voice comparison. University of applied sciences Koblenz, Germany. Cited from: [https://www.kfs.oeaw.ac.at/publications/iafpa\\_abstracts/nr21\\_becker\\_revised.pdf](https://www.kfs.oeaw.ac.at/publications/iafpa_abstracts/nr21_becker_revised.pdf).

Becker, T., Solewicz, Y., Jardine, G. and Gfrorer, S. (2012). Comparing automatic forensic voice comparison systems under forensic conditions. Proceedings of the Audio Engineering Society (AES) International Conference in Audio Forensics, Denver, Colorado, U.S.

Beigi, H. (2011). The fundamentals of speaker recognition. 1<sup>st</sup> Edition. Boston, MA: Springer Publishing, U.S..

Bennyassine, A., Schlomot, E. and Su, H.Y., Massaloux, D., Lamblin, C., Petit, J.P (1997) ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9), pp.64–73.

Beritelli, F. (2008). Effect of background noise on the SNR estimation of biometric parameters in forensic speaker recognition. Proceedings from the Second International Conference on Signal Processing and Communication Systems, pp.1-5.

Beritelli, F., Casale, S., Grasso, R. and Spadaccini, A. (2010). Performance evaluation of SNR estimation methods in forensic speaker recognition. Fourth International Conference on Emerging Security Information, Systems and Technologies, pp.88-92.

Besacier, L., Grassi, S., Dufaux, A., Ansorge, M. and Pellandini, F. (2000). GSM Speech coding and speaker recognition. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2, pp.1085-1088.

Besacier, L. and Bonastre, J. (2000). Sub-band architecture for automatic speaker recognition. *Signal Processing Journal* 80 (7), pp.1245-1259.

Bhattacharya, G., Alam, J. and Kenny, P. (2017). Deep speaker embeddings for short-duration speaker verification. Proceedings of the Interspeech Conference, 2017.

Biswas, S., Rohdin, J. and Shinoda, K. (2015). Autonomous Selection of I-vectors for PLDA Modelling in Speaker Verification. *Speech Communication* (72.C), pp.32-46.

Boies, D., Hebert, M. and Heck, L. (2004). Study on the effect of lexical mismatch in text-dependent speaker verification. The Proceedings of the Odyssey Speaker Recognition Workshop, 2004.

Boucheron, L. and De Leon, P. (2008). On the inversion of mel-frequency cepstral coefficients for

- speech enhancement applications. Proceedings of the International Conference on Signals and Electronic Systems, Kraków, Poland, pp.485-488.
- Bricker, P. and Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America* 40(6), pp.1441-1449.
- Bricker, P., Gnanadesikan, R., Mathews, M., Pruzansky, S., Tukey, P., Wachter, K., and Warner, J. (1971). Statistical Techniques for Talker Identification. *The Bell System Technical Journal* (50.4), pp.1427-454.
- Bridle, J. and Brown, M. (1974). An experimental automatic word recognition system. Joint Speech Research Unit, *JSRU Report No.1003*, Ruislip, U.K.
- Brummer, N. and Van Leeuwen, D.A. (2006). On calibration of language recognition scores. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), pp.1-8.
- Bunge, E. (1976). Automatic speaker recognition by computers. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, Pennsylvania, USA.
- Byrne, C. and Foulkes, P. (2004). The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law*, 11(1), pp.83-173.
- Campbell, J. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, vol. 85(9), pp.1437-1462.
- Campbell, J. (2014). Speaker recognition for forensic applications. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), Joensuu, Finland.
- Campbell, W.M., Sturium, D.E. and Reynolds, D.A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5), pp.308-31.
- Cardoso, A., Foulkes, P., French, J. P., Gully, A. J., Harrison, P. T. and Hughes, V. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. Proceedings of the 19th International Congress of Phonetic Sciences, (ICPhS), <http://eprints.whiterose.ac.uk/142644/>.
- Castellano P.J., Sridharan S. and Cole D. (1996). Speaker recognition in reverberant enclosures. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia (1), pp.117-120.
- Castellanos, A., Benedí, J.M. and Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, 20(1-

2), pp.23-35.

- Chakroun, R., Frikha, M. and Zouari, L.B. (2018). A New Approach for Short Utterance Speaker Identification. *Iet Signal Processing*, (12.7), pp.873-80.
- Champod, C. and Evett. I. W. (2001). A probabilistic approach to fingerprint evidence. *Journal of Forensic Identification*, (51) pp.101-122.
- Chang, J. and Wang, D. (2017). Robust speaker recognition based on DNN/i-vectors and speech separation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5415-5419.
- Chen, N., Shen, W., Campbell, J. and Schwartz, R. (2009). Large-scale analysis of formant frequency estimation variability in conversational telephone speech. *Proceedings of the 10<sup>th</sup> Annual Conference of the International Speech Communication Association (ISCA)*, Brighton, U.K., pp.2203-2206.
- Chen, S., Xu, M. and Pratt, E. (2012). Study on the effects of intrinsic variation using i-vectors in text independent speaker verification. *Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey)*, Singapore, pp.172-179.
- Clark, H.H. and Fox Tree, J.H. (2002). Using ‘uh’ and ‘um’ in spontaneous speaking. *Cognition*, 84(1), pp.73-111.
- Clarke, F.R. (1965). Speaker recognition by humans. *The Journal of the Acoustical Society of America* 37(6), pp.1211-1212.
- Dabbs, J.H. and Schmidt, O.L. (1972). Apollo experience report voice communications techniques and performance. N.A.S.A Technical Archive. Manned space research, Houston, Texas, NASA, Washington D.C.
- Dario, J. and Barbosa, R. (2012). Reverberation: convolution and algorithms. Cited from:  
<https://ses.library.usyd.edu.au/bitstream/handle/2123/8307/InitialTechnologicReviewFinal.pdf?sequence=2&isAllowed=y>
- Das, R.K., Jelil, S. and Prasanna, S.M. (2016). Significance of constraining text in limited data text-independent speaker verification. *Proceedings of SPCOM, 2016*, pp.1-5.
- Das, R.K. and Prasanna, S.R. (2017). Speaker verification from short utterance perspective: A review. *The Institute of Electronics and Telecommunication Engineers (IETE) Technical Review*, 35(6), pp.599-617.

- Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). Cited from: <http://euro.ecom.cmu.edu/program/law/08-732/Evidence/Daubert-Dow.pdf>.
- Decker, J. and Handler, J. (1977). Voiceprint identification evidence-out of the Frye pan and onto admissibility. *The American University Law Review* 26(314), pp.314-372, republished 2010.
- Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4), pp.788-798.
- Deshpande, M.S. and Holambe, R.S. (2011a). Speaker identification based on robust AM-FM features. The 2<sup>nd</sup> International Conference on Emerging Trends in Engineering and Technology (ICETET), pp.880-884.
- Deshpande, M. S. and Holambe, R. S. (2011b). Robust speaker identification in presence of car noise. *International Journal of Biometrics*, 3(3) pp.189-205.
- Doddington, G.R. (1970). A Method of Speaker Verification; PhD Thesis, University, Wisconsin, U.S.A.
- Doddington, G., Martin, A., Kamm, T., Ordowski, M. and Przybocki, M. (1997). The DET curve in assessment of detection task performance. Cited from: <https://apps.dtic.mil/dtic/tr/fulltext/u2/a530509.pdf>
- Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D (1998). Sheep, goats, lambs and wolves: a statistical analysis of speaker performance. Cited from: <https://apps.dtic.mil/dtic/tr/fulltext/u2/a528610.pdf>
- Drygajlo A. (2012). Automatic speaker recognition for forensic case assessment and interpretation. *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. Springer New York, pp.21-39.
- Drygajlo, A., Meuwly, D. and Alexander, A. (2003). Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition. *Proceedings of Eurospeech Switzerland*, pp. 689-692.
- Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen J. and Niemi T. (2015). Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition. *The European Network of Forensic Science Institutes (ENFSI)*. Coted from: [http://enfsi.eu/wp-content/uploads/2016/09/guidelines\\_fasr\\_and\\_fsasr\\_0.pdf](http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf)
- Dulal, S. (2014). Speaker recognition using Gaussian Mixture Models (GMMs). Cited from:

<https://www.slideshare.net/dulalsaurab>

- Dunstone, T. and Yager, N. (2009). *Biometric system and data analysis: Design, evaluation and data mining*. Boston, MA: Springer.
- Ekland, R. (2007). Pulmonic ingressive speech: a neglected universal? *Proceedings of Fonetick* (50), pp.21-24.
- Ekland, R. (2015). Ingressive Phonation and Speech Page. Cited from: <http://ingressivespeech.info/>.
- El-Maleh, K. and Kabal, P. (1997). Comparison of voice activity detection algorithms for wireless personal communications systems. *Canadian Conference on Electrical and Computer Engineering. IEEE. Engineering Innovation: Voyage of Discovery. Conference Proceedings*, vol. 2, pp.470-473.
- ENFSI (2015). *Guidelines for evaluative reporting in forensic science*. Cited from: [http://Enfsi.eu/wp-content/2016/uploads/09/m1\\_guideline.pdf](http://Enfsi.eu/wp-content/2016/uploads/09/m1_guideline.pdf).
- Enzinger, E. (2015). Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence. School of Electrical Engineering & Telecommunications, Faculty of Engineering (PhD Thesis).
- Eriksson, A. (2012). Aural/acoustic methods vs. automatic methods in forensic phonetic casework. In Neustein, A. & Patil, H. A. (eds.). *Forensic speaker recognition: Law enforcement and counter-terrorism*. Berlin: Springer. pp.41-69.
- Esling, J. (2013). Voice and Phonation. Cited from: Jones, M., Knight, R. (eds.). *The Bloomsbury companion to phonetics*. Bloomsbury. London, U.K. pp.110-125.
- ETSI Standards for the GSM Codec (1996). From the European Telecommunications Standards Institute. *Digital cellular telecommunications system (Phase 2+); Mobile Applications*. Cited from: [http://etsi.org/deliver/etsi\\_gts/09/0902/05.03.00\\_60/gsmst\\_0902v050300p.pdf](http://etsi.org/deliver/etsi_gts/09/0902/05.03.00_60/gsmst_0902v050300p.pdf)
- Evans, N., Mason, J., Auckenthaler, R. and Stapart, R. (2002). Assessment of speaker verification degradation due to packet loss in the context of wireless mobile devices. Cited from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.4380>
- Evet, I.W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science and justice* 38(3), pp.198-202.
- Fant, G (1959). *Acoustic analysis and synthesis of speech with applications to Swedish*. Ericsson Technics. Cited from: <https://books.google.co.uk/books?id=aUAhcAAACAAJ>

- Fatima, N. and Zheng, T.F. (2012). Short utterance speaker recognition, a research agenda. *Proceedings of the International Conference on Systems and Informatics (ICSAI)*, pp.1746-1750.
- Fedila, M., Bengherabi, M. and Amrouche, A. (2018). Gammatone filter-bank and symbiotic combination of amplitude and phase-based spectra for robust speaker verification under noisy conditions and compression artefacts. *Multimedia Tools (77)*, pp.16721-1673.
- Ferràs, M., Madikeri, F., Motlicek, P., Dey, S. and Boulard, H. (2016). A large-scale open-source acoustic simulator for speaker recognition. *IEEE Signal Processing Letters* 23(4), pp.527-531.
- Fletcher, H. and Munson, W.A. (1933). Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, (5) pp.82-108.
- French, J.P. (2017). A development history of forensic speaker comparison in the UK. York University and JP French Associates. Cited from <http://eprints.whiterose.ac.uk/117763>.
- French, J.P. and Harrison, P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law* (14), pp.137-144.
- French, J.P., Harrison, P., Cawley, L. and Bhagdin, A. (2009). Evaluation of the Batvox automatic speaker recognition system for use in U.K. based forensic speaker comparison casework. Presented at the *International Association for Forensic Phonetics and Acoustics Annual Conference*. Cambridge, U.K.
- French, J. P., Nolan, F., Foulkes, P., Harrison, P. and McDougall, K. (2010). The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison. *International Journal of Speech Language and the Law* 17(1), pp.143-152.
- French, J.P. and Stevens, L. (2013). Forensic speech science. In Jones, M. and Knight, R.A. *The Bloomsbury companion to phonetics* (12), pp.183-196. London: Continuum.
- Fry, D.B. (1979). *The Physics of Speech*. Cambridge: Cambridge University Press.
- Fujitsu Ltd (2017). Deep learning-based voiceprint authentication from very short speech. *Fujitsu Research & Development Center Co. Ltd*. Beijing: China.
- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. London, UK: Academic Press (1990) 2<sup>nd</sup> Edition.
- Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions*



*on Acoustics, Speech, and Signal Processing* 29(2), pp.254-272.

- Furui, S. (2001). *Digital speech processing, synthesis and recognition*. Marcel Dekker publishing, New York: U.S.
- Gallardo, L.F. (2016). *Human and Automatic Speaker Recognition over Telecommunication Channels*. Singapore: Springer.
- Gallardo, L.F., Wagner, M., and Möller, S. (2012). Analysis of automatic speaker verification performance over different narrowband and wideband telephone channels. Cited from: [http://www.qu.tu-berlin.de/fileadmin/fg41/publications/fernandez-gallardo\\_2012\\_analysis-of-automatic-speaker-verification-performance-over-different-narrowband-and-wideband-telephone-channels.pdf](http://www.qu.tu-berlin.de/fileadmin/fg41/publications/fernandez-gallardo_2012_analysis-of-automatic-speaker-verification-performance-over-different-narrowband-and-wideband-telephone-channels.pdf).
- Ganapathy, S., Pelecanos, J. and Omar, M.K. (2011). Feature Normalization for Speaker Verification in Room Reverberation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4836-839.
- Garcia-Romero, D., Xinhui Z. and Espy-Wilson, C. (2012). Multi-condition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4257-4260.
- Godin, K., Sadjadi, O.S. and Hansen, J. (2013). Impact of noise reduction and spectrum estimation on noise robust speaker identification. Proceedings of the Annual Conference of the International Speech Communication Association (ISCA), pp.3656-3660.
- Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. University of York, Department of Language and Linguistic Science. PhD Thesis.
- Gold, E. and French, J.P. (2011). International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law* 18 (2), pp.293-307.
- Gold, E., French, J.P. and Harrison, P.T. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. Proceedings of the Acoustical Society of America, pp.1-8.
- Gold, E. and French, J.P. (2019). International practices in forensic speaker comparison: second survey. *The International Journal of Speech, Language and the Law* 26 (1) pp.1-20.
- Goldenberg, R., Cohen, A. and Shallom, I. (2006). The Lombard effect's influence on automatic

- speaker verification systems and methods for its comparison. Proceedings of the International Conference on Information Technology: Research and Education, pp. 233-237.
- Gonzalez-Rodriguez, J., Ramos-Castro, D., Garcia Gomar, M. and Ortega- Garcia, J. (2004). On robust estimation of likelihood ratios: the ATVS-UPM system at 2003 NFI/TNO forensic evaluation. Proceedings of Odyssey 2004, the Speaker and Language Recognition Workshop, Toledo, Spain, pp.83-90.
- Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Comar, M. and Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3), pp.331-355.
- Greene and Mathieson's. *The voice & its' disorders*. Sixth addition (2001). London, UK: Wiley and Sons Ltd.
- Guide to the General Data Protection Regulations (GDPR) (2018). Published by the Information Commissioners Office, U.K. Government. Cited from: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/711097/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/711097/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf).
- Gully, A. J., Foulkes, P., French, J.P., Harrison, P. T. and Hughes, V. (2019). Examining articulatory settings using MRI: Pilot results. Proceedings of the International Association of Forensic Phonetics and Acoustics (IAFPA).
- Guo, Jx, Xu, N, Qian, Kl, Shi, Y, Xu, Ky, Wu, Yn, and Alwan, A. (2018). Deep Neural Network Based I-vector Mapping for Speaker Verification Using Short Utterances." *Speech Communication* 105 (2018), pp.92-102.
- Guzewich, P. and Zahorian, S. (2017). Improving speaker verification for reverberant conditions with deep neural network dereverberation processing. Proceedings of the Annual Conference of the International Speech Communication Association (ISCA), Stockholm, Sweden. pp. 171-175.
- Hahn, M.S., Teply, B.A., Stevens, M.M., Zeitels, S.M. and Langer, R. (2006). Collagen composite hydrogels for vocal fold lamina propria restoration. *Biomaterials* (27), pp.1104-1109.
- Hansen, J. and Hasan, T. (2015). Speaker recognition by machines and humans. A tutorial review. *IEEE Signal Processing Magazine*, Nov. '15, pp.74-99.
- Harmse, J., Beck, S. and Nakasone, H. (2006). Speaker recognition score-normalization to compensate for SNR and duration. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), San Juan, Puerto Rico, pp.1-8.

- Harrison, P.T. (2013). *Making accurate formant measurements: an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. University of York, Department of Language and Linguistic Science. PhD Thesis.
- Harrison, P.T. and French, J.P. (2010). Assessing the suitability of Batvox for UK casework (part II). Proceedings of the International Association of Forensic Phonetics and Acoustics (IAFPA), Trier, Germany.
- Hasan, T. and Hansen, J. (2011). A study on universal background model training in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19(7), pp.1890-1899.
- Hasan, T., Saeidi, R., Hansen, J. and Leeuwen, D. (2013). Duration mismatch compensation for i-vector based speaker recognition systems. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7663-7667.
- Hatch, A. and Stolcke, A. (2006). Generalized linear kernels for one-versus-all classification: application to speaker recognition. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, pp.v585-v588.
- Hatch, A., Kajarekar, S. and Stolcke, A. (2006). Within-Class Covariance Normalization for SVM-based Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association (ISCA), Pittsburgh, Pennsylvania, U.S., pp.1471-1474.
- Hautamäki, V., Cheng, Y.C., Rajan, P., Lee, C.H. (2013). Minimax i-vector extractor for short duration speaker verification. Interspeech, 2013, pp.3708-3712.
- Hayakawa, S. and Itakura, F. (1994). Text-dependent speaker recognition using the information in the higher frequency band. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (1), pp.137-140.
- Hebert, M. and Heck, L. (2003). Phonetic class-based speaker verification. Proceedings of EUROSPEECH, pp.1665-1668.
- Hebert, M. (2008). Text-Dependent Speaker Recognition. *The Handbook of Speech Processing*, Springer Publishing, U.K. pp.743-762.
- Higgins, A., Bahler, L., and Porter, J. (1991). Speaker Verification Using Randomized Phrase Prompting. *Digital Signal Processing*, 1, pp.89-106.
- Hollien, H. (1990). *The Acoustics of Crime. The New Science of Forensic Phonetics*. U.S.: Plenum publishing.

- Hong, Q., Li, L., Li, M., Huang, L., Wan, L. and Zhang, J. (2015). Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system. *Proceedings of Interspeech*, 2015, pp.1037-1041.
- Honikman, B. (1964). Articulatory settings. In Abercrombie, D., Fry, D.B., MacCarthy, P.A.D., Scott, N.C. and Trim, J.L.M (eds.). *In honour of Daniel Jones*. London: Longman, pp.73-84.
- Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. University of York, Department of Language and Linguistic Science. PhD Thesis.
- Hughes, V., Wood, S. and Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law* 23(1), pp.99-132.
- Hughes, V., Harrison, P.T., Foulkes, P., French, J.P., Kavanagh, C. and San Segundo, E. (2017a). Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of the Interspeech Conference*, Stockholm, Sweden. Cited from: <http://eprints.whiterose.ac.uk/117386>.
- Hughes, V., Harrison, P.T., Foulkes, P., French, J.P., Kavanagh, C. and San Segundo, E. (2017b). The complementarity of automatic, semi-automatic, and phonetic measures of vocal tract output in forensic voice comparison. *Proceedings of the International Association of Forensic Phonetics and Acoustics (IAFPA)*, Split, Croatia.
- Hughes, V., Harrison, P.T., Foulkes, P., French, J.P., Kavanagh, C. and San Segundo, E. (2018). The individual and the system: Assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proceedings of the Interspeech Conference*, Graz, Austria. Cited from: <http://eprints.whiterose.ac.uk/132139>.
- Hughes, V., Harrison, P.T., Foulkes, P., French, J.P. and Gully, A.J. (2019). Effects of formant settings and channel mismatch on semi-automatic systems in forensic voice comparison. *Proceedings of the 19<sup>th</sup> International Congress of Phonetic Sciences (ICPhS)*. Cited from: <http://eprints.whiteroae.ac.uk/142643>
- Janicki, A. (2012). SVM-Based speaker verification for coded and uncoded speech. *Proceedings of the 20<sup>th</sup> European Signal Processing Conference*. Bucharest, Romania, pp.26-30.
- Janicki, A. and Staroszczyk T. (2011). Speaker recognition from coded speech using support vector machines. Cited from: Habernal, I. and Matoušek, V. (eds.). *Text, speech and dialogue*. Berlin: Springer, pp.298-291.

- Jankowski, C., Kalyanswamy, A., Basson, S. and Spitz, J. (1990). NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *International Conference on Acoustics, Speech, and Signal Processing* (1), pp.109-112.
- Jarina, R., Polacký, J., Počta, P. and Chmulík, M. (2017). Automatic speaker verification on narrowband and wideband lossy coded clean speech. *IET Biometrics* (6.4), pp.276-81.
- Jessen, M. (2003). Review article on Rose 2002: forensic speaker identification. *Forensic Linguistics 10(1), Speech Language and the Law*, pp.138-151.
- Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass* (2/4), pp.671-711. London: Blackwell Publishing Ltd.
- Jessen, M., Köster, O. and Gfroerer, S. (2005). The influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech Language and the Law*, 12(2), pp.174-213.
- Jessen, M. and Becker, T. (2010). Long-term formant distribution as a forensic-phonetic feature. Slides from ASA 2<sup>nd</sup> Pan-American/Iberian meeting on acoustics. Mexico.
- Jessen, M., Alexander A. and Forth O. (2014). Forensic voice comparisons in German with phonetic automatic features using VOCALISE software. Proceedings of the Audio Engineering Society (AES) 54<sup>th</sup> International Conference (London, UK), pp.28-35.
- Jessen, M., Bortlík, J., Schwarz, P. and Solewicz, Y. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic\_eval\_01). *Speech Communication*, 111, pp.22-28.
- Jokic, I., Jokic, S., Gnjatovic, M., Secujski, M. and Delic, V. (2011). The impact of telephone channels on the accuracy of automatic speaker recognition. *Telfor Journal*, 3(2), pp.100-104.
- Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S. and Mason, M.W. (2011). I-vector based speaker recognition on short utterances, The Proceedings of The Interspeech Conference, 2011, pp.2341-2344.
- Kanagasundaram, A., Vogt, R., Dean, D.B. and Sridharan, S. (2012). PLDA based Speaker recognition on short utterances. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey).
- Kanagasundaram, A., Dean, D. and Sridharan, S. (2014). Improving PLDA speaker verification with limited development data, Proceedings of the ICASSP Conference, 2014, pp.1665-1669.
- Kavanagh, C. (2012). New Consonantal Acoustic Parameters for Forensic Speaker Comparison.

University of York, Department of Language and Linguistic Science. PhD Thesis.

- Kelly, F. and Hansen, J.H.L. (2016a). (Zoo plot movement for age). Score-aging calibration for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, (24)12, pp.2414-2424.
- Kelly, F. and Hansen, J.H.L. (2016b). Evaluation and calibration of Lombard effects in speaker verification. Proceedings of the Spoken Language Technology Workshop (SLT, IEEE), pp.205-209.
- Kelly, F., Frohlich, A., Dellwo, V., Forth, O., Kent, S. and Alexander, A. (2019). Evaluation of VOCALISE under conditions reflecting those of a real voice comparison case. *Speech Communication* (112), pp.30-36.
- Kenny, P. and Dumouchel, P. (2004). Disentangling speaker and channel effects in speaker verification. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (1), pp.137-140.
- Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P. (2006). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), pp.1435-1447.
- Kersta, L.G. (1962). Voiceprint identification. *Nature* (196), pp.1253-1257.
- Kim, C. and Stern, R. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. Proceedings of the Annual Interspeech Conference, Brisbane, 2008. pp.2598-2601.
- Kinnunen, T., Karpov, E. and Franti, P. (2006). Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech and Language Processing* 14(1), pp.277-288.
- Kinnunen T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* 52(1), pp.12-40.
- Kirchhübel, C. (2009). The effects of Lombard speech on vowel formant measurements. Special presentation on forensic voice comparison and forensic acoustics (poster). Department of Electronics, University of York, UK. Cited from: <http://pdfs.semanticscholar.org/6146/5d3f31e5fdc4331071f892889610f9874a48.pdf>
- Klatt, D.H. and Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* (87)2, pp.820-857.

- Korany, N. (2013). Speaker identification in reverberant conditions. *The Acoustical Society of America* (19), pp.01-08.
- Kreiman, J. and Gerratt, B. (2000). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustic Society of America*, 108(4), pp.1867-1879.
- Kreiman, J., Vanlancker-Sidtis, D. and Gerratt, B. (2003). Defining and measuring voice quality. Proceedings of the Annual Conference of the International Speech Communication Association (ISCA), Geneva, Switzerland, pp.115-120.
- Kreiman, J., Gerratt, B. and Ito, M. (2007) When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America* 122(4): pp.2354-364.
- Künzel, H. (2001). Beware of the ‘telephone effect’: the influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law*, 8(1), pp.80-99.
- Künzel, H. and Alexander, P. (2014). Forensic automatic speaker recognition with degraded and enhanced speech. *Journal of the Audio Engineering Society*, 62(4), pp.244-253.
- Larcher, A., Lee, K., Li, H., and Bonastre, F. (2012). I-vectors in the Context of Phonetically-constrained Short Utterances for Speaker Verification. Proceedings from the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4773-776.
- Larcher, A., Lee, K., Ma, B. and Li, H. (2013). Phonetically-constrained PLDA Modeling for Text-dependent Speaker Verification with Multiple Short Utterances." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013): 7673-677
- Larcher, A., Bonastre, J.F., Fauve, B., Lee, K., Levy, C., Li, H., Mason, J. and Parfait, J.Y. (2013). ALIZE 3.0-Open Source Toolkit for State-of-the-Art Speaker Recognition. Proceedings from the Annual Conference of the International Speech Communication Association (ISCA), Lyon, France, pp.2768-2772.
- Larcher, A., Lee, K., Ma, B. and Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication*, 60, pp.56-77.
- LaRiviere, C.J., Winitz, H. and Herriman, E. (1975). The distribution of perceptual cues in English prevocalic fricatives. *Journal of Speech and Hearing Research*, 18, pp.613-622.
- Lau, P. (2008). The Lombard effect as a communicative phenomenon. UC Berkeley, annual report, (4), pp.001-009. Cited from <http://escholarship.org/uc/item/19j8j0b6>.

- Laver, J.D.M. (1968). Voice quality and indexical information. *International Journal of Language & Communication Disorders*, 3(1), pp.43-54. London, UK: Wiley and Sons Ltd.
- Laver, J.D.M. (1975). Individual features of voice quality. University of Edinburgh, Scotland. PhD Thesis.
- Laver, J.D.M. (1979). *Voice quality: a classified research bibliography*. Amsterdam, Netherlands: John Benjamin publishing.
- Laver, J.D.M. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
- Laver, J.D.M. (1991). *The gift of speech: papers in the analysis of speech and voice*. Edinburgh, Scotland: Edinburgh University Press.
- Laver, J.D.M., Wirz, S., Mackenzie, J. and Hiller, S. (1981): A perceptual protocol for the analysis of vocal profiles. Edinburgh, University of Edinburgh, pp.265-280.
- Li, N. and Mak, M. (2015). SNR-invariant PLDA modelling in non-parametric subspace for robust speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(10), pp.1648-1659.
- Li, N. and Mak, M. (2016). SNR-invariant PLDA with multiple speaker subspaces. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICSSP), pp.5565-5569.
- Lindh, J. (2004). Handling the “voiceprint” issue. Proceedings of FONETIK 2004, Dept. of Linguistics, Stockholm University. Department of Linguistics Göteborg University.
- Lode, M., Örtl, M., Koch, C., Rizk, A. and Steinmetz, R. (2018). Detection and analysis of content creator collaborations in Youtube videos using face and speaker recognition. Cornell University. Cited from: <https://arxiv.org/pdf/1807.02020.pdf>, pp.01-12.
- Lombard, E. (1911). Le signe de l'élévation de la voi. Published: Ann Malad, l'Oreille Larynx. Volume 37, pp. 101-119. *The sign of the elevation of the voice*, translated by Mason, P. (2006). Cited from: <http://www.paul.sobriquet.net/wp-content/uploads/2007/02/lombard-1911-p-h-mason-2006.pdf>.
- M.I.T. Lincoln Laboratories. (Date not provided). Cohorts and sub populations for imposters and for UBM. Cited from: <http://www.ll.mit.edu/mission/communications/publications>.
- Ma, J., Sethu, V., Ambikairajah, E. and Lee, K. (2017). Duration compensation of i-vectors for short



- duration speaker verification. *Electronics Letters* 53(6), pp.405-407.
- Maclay, H. and Osgood, C. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15(1), pp.19-44.
- Mamodiya, S., Kumar, L., Das, R.K. and Prasanna, S.M. (2016). Exploring acoustic factor analysis for limited test data speaker verification. Proceedings of TENCON, 2016, pp.1397-1401.
- Mandasari, M.I., McLaren, M. and Van Leeuwen, D.A. (2011). Evaluation of i-vector speaker recognition systems for forensic application, Proceedings of Interspeech, 2011, pp.21-24.
- Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W., Taroni, F. and Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science and Justice* 56(5), pp.364-370.
- Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M. (1997). The DET curve in assessment of detection task performance. Technical report cited from DoD/NIST, Gaithersburg.
- Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, 9(5), pp.504-512.
- McDougall, K., Duckworth, M. and Hudson, T. (2015). Individual and group variation in disfluency features: A cross-accent investigation.' Proceedings of the 18th International Congress of Phonetic Sciences, (ICPhS), Glasgow, Scotland. Cited from: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0308.pdf>, pp. 1-5.
- McDougall, K. and Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication* (95), pp.16-27.
- McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, New York, U.S.A.
- Memon, S., Lech, M. and He, L. (2009). Using information theoretic vector quantization for inverted MFCC based speaker verification. Proceedings of the 2009 2<sup>nd</sup> International Conference on Computer, Control and Communication, pp.01-05.
- Mermelstein, P. (1976). Distance Measures for Speech Recognition. Psychological and Instrumental. *Haskins Laboratories Status Report on Speech Research* (47), pp.91-103.

- Mermelstein, P. and Davis, S. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* (28), No. 4, pp.357-366.
- Meuwly, D., Ramos, D. and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International* (276), pp.142-153.
- Ming, J., Hazen, T., Glass, J. and Reynolds, D. (2007). Robust speaker recognition in noisy conditions. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICSSP) 15(5), pp.1711-1723.
- Miro, X.A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE transactions on audio, speech and language processing*, 20(2), pp.1-15.
- Misra, H., Ikbal, S. and Yegnanarayana, B. (2003). Speaker-specific mapping for text-independent speaker recognition. *Speech Communication*, (39), pp.301-310.
- Mitra, V., Franco, H. and Graciarena, M. (2013). Damped Oscillator Cepstral Coefficients for robust speech recognition. Proceedings of the Interspeech Conference, 2013. Cited from: [https://archive.sri.com/sites/default/files/publications/interspeech\\_2013\\_sydocc\\_v.\\_mitra\\_final.pdf](https://archive.sri.com/sites/default/files/publications/interspeech_2013_sydocc_v._mitra_final.pdf), pp.01-05.
- Morrison, G.S. (2009). Typicality, similarity and calculation of LR. Cited From: [acoustics.org/pressroom/httpdocs/157th/morrison.html](http://acoustics.org/pressroom/httpdocs/157th/morrison.html).
- Morrison, G.S. (2018a). Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review*, (1), pp.20-33.
- Morrison, G.S. (2018b). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International* (283), pp.e1-e7.
- Morrison, G.S., Thiruvaran, T. and Epps, J. (2010). Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), Brno, Czech Republic, pp.63-70.
- Morrison, G.S., Ochoa, F. and Thiruvaran, T. (2012). Database selection for forensic voice comparison. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), Singapore, pp.62-77.

- Morrison, G.S., Sahito, F., Jardine, G., Djokic, D., Clavet, S., Berghs, S. and Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* (263), pp.92-100.
- Morrison G.S. and Enzinger E. (2018). Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality. *Science & Justice* (58), pp.47-58.
- Morrison G.S. and Enzinger E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic\_eval\_01*) – conclusion. *Speech Communication*, (112), pp.37-39.
- Mullen, C., Spence, D., Moxey, L. and Jamieson, A. (2014). Perception problems of the verbal scale. *Science and Justice* (54), pp.154-158.
- Nakasone, H. (2003). Automated speaker recognition in real world conditions: controlling the uncontrollable. Proceedings of The Eurospeech Conference, Geneva, pp.697-700.
- Nandan, N. and Saha, G. (2012). On the performance of IP and mobile based Automatic Speaker Verification. Proceedings of the National Conference on Communications (NCC), pp.1-5.
- Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.
- Nolan, F. (2002). The ‘telephone effect’ on formants: a response. *The International Journal of Speech, Language and the Law*, 9(1), pp.74-82.
- Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12(2), pp.143-173.
- Nolan, F. and McDougall, K., De Jong, G. and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, 16(1), pp.31-57.
- O'Connor, K. (2013). *Examination of stability in fingerprint recognition across force levels using zoo plots*. MSc Dissertation, Purdue University. Cited from: <https://pdfs.semanticscholar.org/edfe/80923f92064e57fcad6db95e5ec4bbccce833.pdf>, pp. 01-89.
- Orman, O.D. and Arslan, L.M. (2001). Frequency analysis of speaker identification. Proceedings of The Speaker and Language Recognition Workshop (IEEE Speaker Odyssey), Crete, Greece, pp.01-04.

- Peer, I., Rafaely, B. and Zigel, Y. (2008). Reverberation Matching for Speaker Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing , pp.4829-832.
- Petracca, M., Servetti, A. and DeMartin, J.C. (2006). Performance analysis of compressed-domain automatic speaker recognition as a function of speech coding technique and bit rate. Proceedings of the IEEE International Conference on Multimedia and Expo, pp.1393-1396.
- Pisoni, D. and Remez, R. (Eds.) (2004). *The handbook of speech perception*. London, U.K.: Wiley-Blackwell Publishing.
- Poddar, A., Sahidullah, M. and Saha, G. (2015). Performance comparison of speaker recognition systems in the presence of duration variability. Proceedings of the Annual IEEE India Conference (INDICON), pp.1-6.
- Poddar, A., Sahidullah, M. and Saha, G. (2018). Speaker verification with short utterances: a review of challenges, trends and opportunities. *Institution of Engineering and Technology, Biometrics Journal* 7(2), pp.91-101.
- Pohlmann, K. (2011). *Principles of Digital Audio* (6th Edition). U.S.A: McGraw Hill Publishing.
- Polacký, J., Počta, P. and Jarina, R. (2016a). An impact of wideband speech codec mismatch on a performance of GMM-UBM speaker verification over telecommunication channel. *Proceedings of the IEEE 11th International Conference from ELEKTRO*, pp.77-82.
- Polacký, J., Počta, P. and Jarina, R. (2016b). Influence of packet loss on a speaker verification system over IP network. *26<sup>th</sup> International Conference of Radioelektronika*, pp.390-394.
- Pollack, I., Pickett, J. M. and Sumbly, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, 26, pp.403-406.
- Pradhan, G. and Prasanna, M. (2011). Significance of speaker information in wideband speech. Proceedings of the Annual National Conference on Communications (NCC), pp.1-5.
- Prasanna, M. and Pradhan, G. (2011). Significance of vowel-like regions for speaker verification under degraded conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (8), pp.2552-2565.
- Prince, S.J.D. and Elder, J.H. (2007). Probabilistic linear discriminant analysis for inferences about identity. Proceedings of the IEEE 11th International Conference on Computer Vision, (ICCV), pp.01-08.
- Regina -v- Slade and Ors. [2015]. EWCA, Crim 71. Report 2015, appeal November 2014, conviction February 2010.

- Reynolds, D.A. (1992). A gaussian mixture modeling approach to text-independent speaker identification. PhD thesis. Georgia institute of technology.
- Reynolds, D.A. (1994). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication* (17,1-2), pp.98-108.
- Reynolds, D.A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C. and Carlson, B. A. (1995). The effects of telephone transmission degradations on speaker recognition performance. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1), pp.329-332.
- Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. (2000). Speaker verification using adapted gaussian Mixture Models. *Digital Signal Processing* (10), pp.19-41.
- Reynolds, D.A. and Sturim, D. (2005). A speaker adaptive cohort selection for T-norm in the text dependent speaker verification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1), pp.1741-1744.
- Rose, P. (2002). *Forensic Speaker Identification*. London and New York: Taylor and Francis.
- Rose, P. (2013). More is better. *The International Journal of Speech, Language and the Law* 20(1), pp.77-116.
- Rose, P. and Morrison, G.S. (2009). A response to the UK Position Statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law* 16(1), pp.139-163.
- Rosenberg, A.E. (1976). Automatic speaker verification: A review. *Proceedings of the 1976 IEEE International Conference*, 64(4), pp.475-487.
- Rosenberg, A.E., Lee, C.H. and Soong, F. (1990). Sub-word unit talker verification using hidden Markov models. *IEEE International Conference on Acoustics, Speech and Signal Processing* (1), pp.269-272.
- Sadjadi, O.S. and Hansen, J. (2010). Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. *Proceedings of the Interspeech Conference*, Makuhari, Japan, pp.2138-2141.
- San Segundo, E.F. and Mompean, A. (2017) Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice* (15,2), pp.1-17.
- San Segundo, E.F., Foulkes, P., French, J.P., Harrison, P.T., Hughes, V. and Kavanagh, C. (2018). The use of the vocal profile analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association*, (49,3), pp.353-380.

- Saquib, Z. Nirmla, S., Rekha, N.P., Nipun, P., Alanksha Kin, J., Tai-Hoon, Sankar, P.K., Grosky, W., Pissinou, N., Shih, T.K. and Slezak, D. (2010). A Survey on automatic speaker recognition systems. Proceedings of the Signal Processing and Multimedia: International Conferences, SIP and MulGrab, (FGIT), pp.134-145. Berlin, Heidelberg: Springer publishing.
- Sarkar, A., Matrouf, D., Bousquet, P. and Bonastre, J. F. (2012). The study of the effect of i-vector modelling on short and mismatch utterance duration for speaker verification. Proceedings of the 2012 Interspeech Conference, Portland, U.S., pp.2662-2665.
- Saunders, J. (1996). Real-time discrimination of broadcast speech/music. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2, pp.993-996.
- Scheffer, N., Luciana, F., Lawson, A.D., Lei, Y. and McLaren, M. (2013). Recent developments in voice biometrics: Robustness and high accuracy. Proceedings of the 2013 IEEE International Conference on Technologies for Homeland Security (HST), pp.447-452.
- Scheffer, N. and Lei, Y. (2014). Content matching for short duration speaker recognition. Proceedings of the Interspeech Conference, Singapore, pp.1317-1321.
- Scherer, K.R. and Giles, H. (1980). *European studies in social psychology*. New York: Cambridge University Press.
- Schiel, F. and Zitzelsberger, T. (2018). Evaluation of formant trackers. *The Bavarian archive for speech signals*, pp.2843-2848. Cited from: <https://www.phonetik.uni-muenchen.de/forschung/publikationen/ZitzelsbergerSchiel-LREC2018-28.pdf>.
- Schnitzer, D., Flexer, A. and Schluter, J. (2013). The relation of hubs to the Doddington zoo in speaker verification. Proceedings of the 21<sup>st</sup> European Signal Processing Conference (EUSIPCO), pp.1-5.
- Schroeder, M. (1964). A new method of measuring reverberation time. Bell telephone research laboratories, Murray Hill, New Jersey, U.S.A. Report from December 1964.
- Schwartz, J.C., Whyte, A.T., Al-Nuami, M. and Donai, J.J. (2018). Effects of signal bandwidth and noise on individual speaker identification. *The Journal of the Acoustical Society of America*, 144(5), pp.EL447-EL452.
- Schwartz, R., Campbell, J. and Shen, W. (2011). When to punt on speaker comparison? *The Journal of the Acoustical Society of America*, 130(4), pp. 2547-2548.
- Shabtai, N., Rafaely, B. and Zigel, Y. (2010). The effect of reverberation on optimal GMM order

and CMS performance in speaker verification systems. *Advances in speaker recognition*.  
Cited from: <https://cdn.intechopen.com/pdfs/11854.pdf>.

Shaver, C.D. and Acken, J.M. (2016). A brief review of speaker recognition technology. Cited from:  
Portland state university online at <https://pdxscholar.library.pdx.edu>

Silovsky, J., Cerva P., and Zdansky, J. (2011). Assessment of speaker recognition on lossy codecs  
used for transmission of speech. *Proceedings of ELMAR, The 53<sup>rd</sup> International Symposium  
Electronics in Marine, Zadar, Croatia*, pp.205-208.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018). X-Vectors: Robust  
DNN embeddings for speaker recognition. *IEEE International Conference on Acoustics,  
Speech and Signal Processing*, pp.5329-5333.

Sohn, J., Kim, N.S. and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE  
Signal Processing Letters*, 6(1), pp.1-3.

Solewicz, Y.A, Becker, T., Jardine, G. and Gfroerer, S. (2012). Comparison of speaker recognition  
systems on a real forensic benchmark. *Proceedings of The Speaker and Language  
Recognition Workshop (IEEE Speaker Odyssey), Singapore*, pp.86-91.

Solewicz, Y.A., Jessen, M. and Van Der Vloed, D. (2017). Null-hypothesis LLR: A proposal for  
forensic automatic speaker recognition. *Proceedings of the Interspeech Conference,  
Stockholm, Sweden*, pp.2849-2853.

Stauffer, A.R. and Lawson, A.D. (2009). Speaker recognition on lossy compressed speech using the  
Speex codec. *Proceedings of the Interspeech Conference, Brighton, UK*, pp.2363-2366.

Stevens, L. and French, J.P. (2012). Voice quality in studio quality and telephone transmitted  
recordings. *The British Association of Academic Phoneticians (BAAP) Conference, Leeds,  
March 2012*.

Stevens, S., Volkman, J. and Newman, E. (1937). A scale for the measurement of the psychological  
magnitude pitch. *The Journal of the Acoustical Society of America*, (8), pp.185-190.

Sturim, D., Campbell, W., Reynolds, D., Dunn, R. and Quatieri, T.F. (2006). Robust speaker  
recognition with cross channel data. MITLL Results on the 2006 NIST SRE Auxiliary  
Microphone Task. *Proceedings of the International Conference on Acoustics, Speech and  
Signal Processing (ICASSP) (4)*, pp.(iv)49-(iv)52.

Styler, D. (2017). *Save the vowels, using Praat for linguistic research*. Cited from: [http://  
savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf](http://savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf).

- Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. and Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3), pp.917-928.
- Tirumalaa, S., Sremath, S., Shahamiri, S., Garhwal, A. and Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications* (90), pp.250-271.
- Togneri, R. and Pullella, D. (2011). An overview of speaker identification: accuracy and robustness issues. *The IEEE Circuits and Systems Magazine* 11(2), pp.23-61.
- Tranter, S. and Reynolds, D. (2006). An Overview of Automatic Speaker Diarization Systems. *IEEE Transactions on Audio, Speech, and Language Processing* 14.5 (2006): pp.1557-565.
- Vanderslice, R. and Ladefoged, P. (1967). The 'voiceprint' mystique. *The Journal of the Acoustical Society of America*, 42(5), p.1164.
- Voiers, W.D. (1961). Perceptual Criteria of Speaker Identity. *The Journal of the Acoustical Society of America*, 33(11), pp.1677-1678.
- Voiers, W.D. (1964). Perceptual Bases of Speaker Identity. *The Journal of the Acoustical Society of America*, 36(6), pp.1065-1073.
- Wan, V. and Renals, S. (2005). Speaker verification using sequence discriminant support vector machines. *IEEE Transactions on Speech and Audio Processing*, 13(2), pp.203-210.
- Wang, X., Hughes, V. and Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech, Language and the Law* (Publishing Summer 2020). Draft cited from <https://www.researchgate.net/journal/1748-8885>.
- Watt, D. and Burns, J. (2012). Verbal descriptions of voice quality differences among untrained listeners. *University of York, Papers in Linguistics*, (2), pp.1-28.
- Waves Software Guides (2014-2019). A product manual for Waves IR-L reverberation. Cited from: <http://waves.com/plugins/ir-l-convolution-reverb>.
- Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustic Society of America*, 51(6ii), pp.2044-2056.
- Wu, B., Li, K., Yang, M. and Lee, C. (2017). A Reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech*



*and Language Processing*, 25(1), pp.102-111.

- Wu, Z., Yamagishi, J., Kinnunen, T., Hanilc, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M. and Delgado, H. (2017). ASR spoof: The automatic speaker verification spoofing and countermeasures challenge. *The IEEE journal of selected topics in signal processing*, 11(4).
- Xue, S and Hao, J. (2005). Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *The Journal of Voice*, 20(3), pp.391-400.
- Yager N. and Dunstone T. (2010). *The biometric menagerie*. IEEE Transactions on pattern analysis and machine intelligence, 32(2), pp.220-230.
- Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T. and Kellermann, W. (2012). Making machines understand us in reverberant rooms. *IEEE: Robustness against reverberation for automatic speech recognition*, 29(6), pp.114-126.
- Zhang, C. and Tang, C. (2018). Evaluation of Batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (*forensic\_eval\_01*). *Speech Communication*, (100), pp.13-17.
- Zhao, X., Wang, Y. and Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on audio, speech and language processing*, 22(4), pp.836-845.
- Zieger, C. and Omologo, M. (2008). Combination of clean and contaminated GMM/SVM for far-field text-independent speaker verification. Proceedings of the Interspeech Conference, Brisbane, Australia, pp.1949-1952.
- Zimmerman trial (2013). Cited from: <http://legalinsurrection.com/2013/06/zimmerman-prosecutions-voice-expert-admits-this-is-not-really-good-evidence>.

**“Better to be despised for too anxious apprehensions  
than ruined by too confident security”**

**Edmund Burke, British Philosopher  
(1723-1792)**