RMIT UNIVERSITY

Automatic Emotion Recognition in Noisy, Coded and Narrow-Band Speech

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

Abas Albahri

M.Eng.Control System and Measurement
B.Eng. Electrical and Electronic

School of Engineering
College of Science, Engineering and Health
RMIT University

September 2016

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Abas Albahri

August 2016

Table of Contents

Declaration ..... **ii**

Table of Contents ..... **iii**

List of Tables ..... **vii**

List of Figures ..... **viii**

Acknowledgements ..... **x**

Abstract ..... **xi**

List of Publications ..... **xiii**

List of Abbreviations ..... **xiv**

Chapter 1: Introduction **1**
1.1 Preview 1
1.2 Background and Problem Statement 1
1.3 Thesis Aims 2
1.4 Thesis Scope 2
1.5 Thesis Research Questions 3
1.6 Thesis Major Findings and Contributions 4
1.7 Thesis Narrative and Structure 5
1.7.1 Thesis Narrative 5
1.7.2 Thesis Structure 5

Chapter 2: Literature Review of AER **8**
2.1 Preview 8
2.2 Automatic Emotion Recognition (AER) from Clean Full Band Speech 9
2.2.1 AER from images and video sequences 9
2.2.2 Advantages of AER from speech signals 9
2.2.3 Challenges of data collection for AER from speech signals 10
2.2.4 Methodology of AER from speech signals 11
2.2.5 Speech feature parameters for AER 12
2.2.6 Long and short term approaches to speech feature extraction for AER 13
2.2.7 AER approach used in the thesis 16
2.3 Automatic Emotion Recognition (AER) from Noisy and Limited Band Speech 16
2.3.1 AER from noisy speech 17
2.3.2 Effects of telephone band reduction and speech coding on speech signals 19
2.4 Conclusion 24

Chapter 3: Effect of Speech Compression on AER **26**
3.1 Preview 26
3.2 Introduction 26
3.3 Advantages of Speech Compression Introduced to Industry 26
3.4 Possible Factors Associated with Speech Compression 27
3.5 Chapter Aims 27
3.6 Method 28
3.6.1 Speech Database 28

3.6.2 Experimental Framework....28
3.6.3 Average Percentage of Identification Accuracy....29
3.6.4 Speech Compression Methods....29
3.6.5 Speech Features....31
3.6.6 Modelling and Classification Methods....33
3.7 Results and Discussion....33
3.7.1 Classification Outcomes for Uncompressed Speech....33
3.7.2 Effect of Narrow-Band Adaptive Multi-Rate AMR Compression on Emotion Classification....34
3.7.3 Effect of AMR-WB Compression on Emotion Classification....35
3.7.4 Effect of AMR-WB+ Compression on Emotion Classification....36
3.7.5 Effect of MP3 Compression on Emotion Classification....36
3.8 Conclusion....38

Chapter 4: Effect of Band Reduction on AER....42
4.1 Preview....42
4.2 Method....43
4.2.1 Speech Data....43
4.2.2 Experimental Framework....43
4.2.3 Bandwidth Reduction....45
4.2.4 Calculation of Speech Features for Emotion Recognition....45
4.3 Experiments and Results....47
4.4 Conclusion....48

Chapter 5: Effect of Speech Compression and Hearing Loss Simulation on AER....50
5.1 Preview....50
5.2 Hearing Loss....50
5.3 Method....51
5.3.1 Database....51
5.3.2 Experimental Framework....51
5.3.3 Hearing Loss Simulation....53
5.3.4 Calculation of Speech Features for Emotion Recognition....54
5.4 Results and Discussion....55
5.4.1 Effect of AMR and Hearing Loss Simulation on AER....56
5.4.2 Effect of AMR-WB and Hearing Loss Simulation on AER....59
5.4.3 Effect of AMR-WB+ and Hearing Loss Simulation on AER....62
5.5 Conclusion....64

Chapter 6: Effect of Speech Compression and Noise on AER....66
6.1 Preview....66
6.2 Method....66
6.2.1 Database....66
6.2.2 Experimental Framework....67
6.2.3 Noise Addition....68
6.2.4 Calculation of Speech Features for Emotion Recognition....69
6.3 Discussion and Results....70
6.3.1 Effect of 15dB Noisy Speech with Three Different Standard Speech Compression Techniques (AMR, AMR-WB and AMR-WB+)....70
6.3.2 Effect of 10dB Noisy Speech with Three Different Standard Speech Compression Techniques (AMR, AMR-WB and AMR-WB+)....73

6.3.3 Effect of 5dB Noisy Speech with Three Different Standard Speech Compression Techniques....76
6.4 Conclusion....79

Chapter 7: Effect of Speech Compression on AER Based on Speech Spectrograms....81
7.1 Preview....81
7.2 Introduction....81
7.3 Method....82
7.3.1 Speech Data....82
7.3.2 Speech Spectrograms....82
7.3.3 Calculation of Speech Spectrograms (SS)....83
7.3.4 Calculation of Speech Spectrogram Energy Features Derived from Critical Bands (SS-CB) and Bark Bands (SS-Bark)....84
Critical Bands[Hz]....86
7.3.5 Speech Compression Methods....87
7.4 Results and Discussion....88
7.4.1 Classification Outcomes for Uncompressed Speech....88
7.4.2 Effect of the Narrow-Band AMR Compression on Emotion Classification....88
7.4.3 Effect of AMR-WB Compression on Emotion Classification....89
7.4.4 Effect of AMR-WB+ Compression on Emotion Classification....92
7.4.5 Effect of MP3 Speech Compression on Emotion Classification....93
7.5 Conclusion....94

Chapter 8: Artificial Bandwidth Extension to Improve AER from Narrow-Band Coded Speech....96
8.1 Preview....96
8.2 Introduction....96
8.3 Method....97
8.3.1 Narrow-Band Speech Compression....97
8.3.2 Artificial Bandwidth Extension....99
8.3.3 Signal Up-Sampling....99
8.3.4 Bandwidth Extension Using Source-Filter Model....99
8.3.5 Bandwidth Extension Using Spectral Folding....100
8.3.6 Removal of Processing Artefacts....101
8.4 Experiments and Results....101
8.4.1 Conversational Data....101
8.4.2 Speech Emotion Recognition....101
8.5 Observations Based on Graphs Shown in Figure 8.3....104
8.5.1 Effect of ABE on AER Results for Different Speech Features....104
8.5.2 Effect of ABE on AER Results for Different Genders....104
8.5.3 Effect of ABE on AER Results for Different AMR Compression Rates....104
8.5.4 AER from Compressed and Uncompressed Speech....105
8.6 Conclusion....105

Chapter 9: Discussion and Summary of Findings....107
9.1 Summary....107
9.2 Discussion and Major Findings....107
9.2.1 What is the Effect of Speech Compression Techniques on AER?....107
9.2.2 What is the Effect of Band Reduction and Coding on AER?....109

9.2.3 What is the Effect of Speech Compression and Hearing Loss Simulation on AER? 109
9.2.4 What are the Combined Effects of Speech Compression and Noise on AER? 110
9.2.5 Is it Possible to Mitigate Detrimental Effects of Speech Compression on AER Using Speech Spectrogram Features? 112
9.2.6 Is it Possible to Reduce Detrimental Effects of Narrow-Band Speech Compression on AER Using Artificial Bandwidth Extension (ABE) of Speech Signals? 113
9.3 Future Work 114
References 116

List of Tables

Table 3.1: Description of the Berlin Emotional Database 28
Table 3.2: Bit rates used in the experiments for different speech compression systems 31
Table 4.1: Frequency bands used in emotion recognition tasks 45
Table 7.1: Critical and Bark Bands 86

List of Figures

Figure 2.1 Training data collection and labelling. 11
Figure 2.2: Standard machine learning approach used in AER. 12
Figure 2.3: Low level speech parameters derived from voiced speech, i.e. speech produced by vocal folds vibration..... 13
Figure 2.4 Long-time, turn based (static) processing of speech feature extraction. 14
Figure 2.5 Short-time, frame based (dynamic) processing of speech feature extraction..... 14
Figure 3.1: Experimental framework effect of compressed speech on SER 29
Figure 3.2: Average accuracy of multi-class emotion recognition for male and female speakers using MFCC, TEO-PWP and GP-T&GP-F features;..... 37
Figure 4.1: Experimental framework for testing effects of band reduction of speech on AER..... 45
Figure 4.2: AER using limited-band speech. 47
Figure 5.1: Experimental framework investigating effects of speech compression and hearing loss simulation on AER..... 52
Figure 5.2: Audiogram setup used to generate a low-pass filter simulating a typical mild-to-moderate hearing loss. The audiogram and the corresponding low-pass filter were generated using publically available AngelSim software http://www.tigerspeech.com/angelsim/angelsim_about.html 54
Figure 5.3: Average accuracy of multi-class emotion recognition for male and female speakers using MFCC features; 56
Figure 5.4: Average accuracy of multi-class emotion recognition for male and female speakers using TEO-PWP features; 57
Figure 5.5: Average accuracy of multi-class emotion recognition for male and female speakers using GP-T&GP-F features;..... 59
Figure 6.1: Emotion speech recognition from noisy compressed or uncompressed speech. White Gaussian noise was added to either compressed or uncompressed speech to generate noisy signals with SNR values of 5dB, 10dB and 15dB..... 68

Figure 6.2: Average accuracy of multi-class emotion recognition for MFCC features with 15dB noise for females and male; 72
Figure 6.3: Average accuracy of multi-class emotion recognition for TEO-PWP features with 15 dB noise for females and males; 73
Figure 6.4: Average accuracy of multi-class emotion recognition for (GP-T, GP-F) features with 15 dB noise for females and males; 73
Figure 6.5: Average accuracy of multi-class emotion recognition for MFCC features with 10 dB noise for females and males; 75
Figure 6.6: Average accuracy of multi-class emotion recognition for TEO-PWP features with 10 dB noise for females and males; 75
Figure 6.7: Average accuracy of multi-class emotion recognition for (GP-T, GP-F) features with 10 dB noise for females and males; 76
Figure 6.8: Average accuracy of multi-class emotion recognition for MFCC features with 5 dB noise for females and males; 78
Figure 6.9: Average accuracy of multi-class emotion recognition for TEO-PWP features with 5 dB noise for females and males; 78
Figure 6.10: Average accuracy of multi-class emotion recognition for GP-T&GP-F features with 5 dB noise for females and males; 78
Figure 7.1: Features generation from the auditory frequency bands of spectrograms using different auditory scales (SS-CB, SS-Bark)..... 83
Figure 7.2: Framework of AER experiments using spectrogram features..... 87
Figure 7.3: Average accuracy of multi-class emotion recognition for males and females using SS-image features; 91
Figure 7.4: Average accuracy of multi-class emotion recognition for males and females using SS-CB features;..... 91
Figure 7.5: Average accuracy of multi-class emotion recognition for males and females using SS-Bark features; 92
Figure 8.1: Bandwidth extension procedure. 98
Figure 8.2: Experimental framework;..... 100
Figure 8.3: Achieved APIA values for uncompressed full-band speech (Un), compressed speech with eight different narrow-band AMR compression rates R1–R8, and for the extended-bandwidth speech (ABE) generated from AMR compressed speech with the same eight different bit rates R1–R8..... 103

Acknowledgements

My words fail to describe the love, fulfilment and appreciation in my heart for the people listed here. I cannot translate my feeling of thanks for them in just a few words. I am grateful to them for their advice and guidance.

I would like to thank my main supervisor, Dr Margaret Lech, for being supportive and helpful during this research project. First, I would like to thank you for giving me this opportunity and trusting me to work with you on this project. Second, thank you for your ideas and your experience, which was a great support to me. Also, thank you for your continued hard work and patience with me during every stage of my study.

I would also like to thank my second supervisor, Elena Pirogova, for her continuing support during my study. Your kind advice was helpful during my candidature.

Special thanks are due to my parents, who worked hard to raise me. Thank you for your love and support throughout my life journey.

Abstract

This thesis addresses an important research gap regarding effects of real-life conditions including coded, narrow-band and noisy speech signals on automatic emotion recognition (AER) from speech signals. In addition, the study aims to research efficient methods of reducing possible detrimental effects of speech signals compression on AER.

The thesis consists of two parts. The first part investigates the effects of noise, data compression and bandwidth reduction on AER from speech signals. The second part investigates application of AER based on speech spectrograms (SS) and the Artificial Bandwidth Extension (ABE) to improve the robustness and accuracy of emotion recognition from speech signals under these potentially undesirable conditions.

Effects of adaptive multi-rates (AMR), adaptive multi-rate wideband (AMR-WB) and extended adaptive multi-rate wideband (AMR-WB+) and MP3 speech compression methods are compared against emotion recognition from uncompressed speech.

Noisy conditions are simulated using Gaussian white noise added to speech signals at different values of signal to noise ratio (SNR). Band reduction is tested using speech filtering.

The AER methods include techniques based on acoustic speech parameters including: mel-frequency cepstral coefficients (MFCCs), Teager energy operator and perceptual wavelet packet (TEO-PWP) features, glottal time and frequency domain features (GP-T&GP-F), as well as, spectrogram image (SS) parameters, spectrogram critical band scale (SS-CB) and spectrogram bark scale (SS-Bark). The modelling of acoustic classes is based on the Gaussian Mixture Mode (GMM) and all experiments use the same Berlin Emotional Speech database.

The ABE of narrow band speech is performed using spectral folding and spectral envelope estimation methods.

The major findings described in this thesis indicate that:

1. Standard speech compression methods such as AMR, AMR-WB, AMR-WB+ and MP3 have a significant effect on the (AER), and in general lead to significant degradation of AER accuracy.

2. Low-frequency components (0 kHz to 1 kHz) of speech containing the fundamental frequency information, as well as, high-frequency components (above 4 kHz) have a key effect on the accuracy of SER.

3. Significant reduction of AER accuracy was observed for uncompressed speech modified in a way simulating a typical mild-to-moderate high frequency hearing loss. This accuracy was further reduced when the modified speech was compressed.

4. Addition of noise to either uncompressed or compressed speech reduces accuracy of AER. It was shown that the best performing under noisy conditions features were MFCCs and the best performing speech compression algorithms was AMR-WB.

5. Detrimental effects of speech compression can be mitigated using AER based on speech spectrogram features.

6. By extending the narrow-band of AMR-compressed speech an improvement of AER accuracy can be achieved.

List of Publications

Journal Publications

1. Albahri A, Lech M, and Cheng E, "Effect of speech compression on the automatic recognition of emotions", *International Journal of Signal Processing Systems*, vol. 4, no. 1, pp. 55–61, 2016 (IJSPS, ISSN: 2315-4535).

Conference Publications

2. Albahri A, Rodriguez C, and Lech M, "Artificial bandwidth extension to improve automatic emotion recognition from narrow-band coded speech", *The 10th IEEE International Conference on Signal Processing and Communication Systems, ICSPCS'2016*, 19-21 December 2016, Surfers Paradise, Gold Coast, Australia, pp 1-6.

3. Albahri A, and Lech M, "Effects of band reduction and coding on speech emotion recognition", *The 10th IEEE International Conference on Signal Processing and Communication Systems, ICSPCS'2016*, 19-21 December 2016, Surfers Paradise, Gold Coast, Australia, pp 1-6.

4. Albahri A, Lech M, and Cheng E, "Effect of speech compression on the automatic recognition of emotions", *International Conference on Signal Processing Systems, ICSPS 2014*, Dubai, 7–10 December 2014.

List of Abbreviations

ABE	Artificial bandwidth extension
ACELP	Algebraic code-excited linear predictive
AER	Automatic emotion recognition
AM	Amplitude modulated
AMR	Adaptive multi-rates
AMR-WB	Adaptive multi-rate wideband
AMR-WB+	Extended adaptive multi-rate wideband
CELP	Code-excited linear prediction
DNN	Deep neural network
GMM	Gaussian mixture model
GP-T&GP-F	Glottal time and frequency domain features
GSM	Global system for mobile communications
HL	Hearing loss
HLS	Hearing loss simulation
HMM	Hidden Markov model
LPC	Linear predictive models
LTASS	Long-term average speech spectrum
LTF	Long-term formant distribution
MFCCs	Mel-frequency cepstral coefficients
MP3	MPEG-1 layer-3 audio

PCM	Pulse code modulation
SER	Speech emotion recognition
SI	Spectrograms image parameters
SNR	Signal-to-noise ratio
SS	Speech spectrogram
SS-Bark	Spectrogram energy features based on bark scale
SS-CB	Spectrograms energy features based on critical band scale
TEO	Teager energy operator parameters
TEO-PWP	Teager energy operator and perceptual wavelet packet
WB	Wideband
HTK	Hidden Markov Model Toolkit
3G WCDMA	Third Generation Wide Band Code Division Multiple Access
PCA	Principal Component Analysis
mRMR	Minimum Redundancy Maximum Relevance
TKK aparat	an environment for voice inverse filtering and parameterization

Chapter 1: Introduction

1.1 Preview

This chapter outlines the problem statement, aims, major funding, contributions and thesis structure.

1.2 Background and Problem Statement

AER from speech signals has many applications, such as speaker recognition, verification and security purposes, and medical and physiological services. However, it is affected by new speech technology applications such as speech compression techniques. Speech compression has many industrial advantages for telecommunications and speech technology, which support and serve speech recognition for human-to-machine communications. These advantages include a reduction in the delay of data transmission using telephony, a reduction in the memory size needed to save speech recordings, and the memory of mobile phones. This results in reduced costs and time saved [38], [19].

Speech compression algorithms can affect performance of speech recognition and classification systems. The coding and decoding procedures can significantly modify temporal and spectral characteristics of speech and change affect both linguistic and paralinguistic (emotional) information. These can directly affect the accuracy of speech recognition [19], [38].

Many previous studies have focused on investigating effects of the speech compression algorithm on speech features used in speech and speaker recognition. These studies indicated that the CELP and LP-based GSM speech codecs have negatively influenced the fundamental frequency (F0) [38], [1]. Further, F0 extracted from the mobile phone speech compression algorithm significantly increases to 30 Hz comparable with F0 extracted from the landline [2], [3]. In addition, the vowel F1 formants extracted from compressed speech using the mobile phone speech compression algorithm is higher than F1 formants extracted from uncompressed speech [3]. Compared with the previous study, F0 and formant frequencies (F1–F3)

decrease significantly when compressed using the GSM AMR speech codec [4]. Speech recognition accuracy has been improved using speech features such as cepstral coefficients compressed by GSM speech codec compared with uncompressed speech [5], [6].

The effects of speech compression on the accuracy of AER from speech signals have not been extensively investigated.

Automatic emotion recognition (AER) from speech signals performed under real-life conditions is likely to deal with coded, narrow-band and noisy speech signals.

Therefore in addition, to investigating effects of speech compression, the thesis aims to research combined effects of noise and speech compression on AER.

The study consists of two parts. The first part investigates the effects of noise, data compression and bandwidth reduction on AER from speech signals. The second part investigates speech-processing techniques that improve the robustness and accuracy of emotion recognition from speech signals under these potentially undesirable conditions.

1.3 Thesis Aims

The thesis addresses an important knowledge gap of how AER from speech signals performs under real-life conditions with coded, narrow-band and noisy speech.

Two thesis aims are:

1. To investigate effects of factors such as speech coding, noise and band reduction on automatic emotion recognition from speech signals;

2. To investigate efficient methods of reducing possible detrimental effects of coding and noise on the AER from speech signals.

1.4 Thesis Scope

The scope is limited to the following research data and methods:

1. All AER algorithms are consistently tested using the same Berlin Emotional Speech (BES) data;

2. In all experiments, the AER framework is consistently setup to perform a simultaneous classification of 7 different categorical emotions: anger, happiness, sadness, fear, disgust, boredom and neutral speech

3. The following speech coding techniques are investigated: AMR, AMR-WB, AMR-WB+ and MP3;

4. The following acoustic speech parameters are used to perform AER experiments: Mel Frequency Cepstral Coefficients (MFCCs), TEO-PWP parameters, glottal time-domain parameters (GP-T) and glottal frequency domain parameters (GP-F);

5. The modelling and classification part of AER is performed using the Gaussian Mixture Model method (GMM).

1.5 Thesis Research Questions

This research provided answers to the following research questions

1. What is the effect of speech compression techniques on AER?

2. What is the effect of band reduction and coding on AER?

3. What are the combined effects of speech compression and hearing loss simulation on AER?

4. What are the combined effects of speech compression and noise on AER?

5. Is it possible to mitigate detrimental effects of speech compression on AER using speech spectrogram features?

6. Is it possible to reduce detrimental effects of narrow-band speech compression on AER using Artificial Bandwidth Extension (ABE) of speech signals?

1.6 Thesis Major Findings and Contributions

Major findings and contributions presented in this thesis can be summarised as follows:

1. It shown that standard speech compression methods such as AMR, AMR-WB, AMR-WB+ and MP3 have a significant effect on the AER, and in general lead to significant degradation of AER accuracy.

2. It was shown that low-frequency components (0 kHz to 1 kHz) of speech containing the fundamental frequency information, as well as, high-frequency components (above 4 kHz) have a key effect on the accuracy of SER. This observation could explain relatively high performance of the wide-band AMR-WB compression. It could also indicate that hearing impairment characterised by loss of high frequencies could lead to reduced ability of understanding speech emotions.

3. It was shown that band reduction of speech signal that simulates a typical high frequency hearing loss has a significant impact on AER and reduces AER accuracy.

4. Significant reduction of AER accuracy was observed for uncompressed speech modified in a way simulating a typical mild-to-moderate high frequency hearing loss. This accuracy was further reduced when the modified speech was compressed.

5. Experimental observations have shown that addition of noise to either uncompressed or compressed speech reduce accuracy of AER. It was shown that the best performing under noisy conditions features were MFCCs and the best performing speech compression algorithms was AMR-WB (most likely due to preservation of wide-band)

6. It was shown that detrimental effects of speech compression can be mitigated using AER based on speech spectrogram features. Speech features representing full spectrogram (SS) have shown only very small (5%) degradation of AER accuracy when applied to speech compressed with AMR-WB and AMR-WB+.

7. The results have shown that by extending the narrow-band of AMR-compressed speech an improvement of AER accuracy can be achieved. The SER results showed that application of artificial bandwidth extension (ABE) lead to at least 5% improvement in emotion recognition accuracy compared to narrow-band compressed speech.

1.7 Thesis Narrative and Structure

1.7.1 Thesis Narrative

After presenting existing body of knowledge in the research areas related to this thesis in Chapter 2, an investigation described in Chapter 3 determines effects of speech compression on AER. Degradation observed in Chapter 3 due to compression can be explained by band limitation introduced by speech compression methods a set of experiments of AER based reduced band was conducted in Chapter 4 and in Chapter 5, where the band limitation simulating a typical age related high frequency hearing loss was investigated in relation to AER. Since real life conditions of speech processing include addition of noise, the combined effects of noise and speech compression on the AER were investigated in Chapter 6. The two final chapters (Chapter 7 and Chapter 8) investigate methods that could potentially mitigate the detrimental effects of speech compression on AER. Thus, Chapter 7 investigates if the use of state of the art speech spectrogram features incorporating auditory perception criteria can improve AER from compressed speech and Chapter 8 applies an Artificial Bandwidth Extension (ABE) to the high end of compressed speech spectrum to determine if an improvement of AER can be achieved.

1.7.2 Thesis Structure

Chapter One outlines the problem statement, aims, major funding, contributions and thesis structure.

Chapter Two provides a literature review on topics related to the thesis. A brief review presented of existing studies related to emotion recognition from speech using acoustic speech features. Different types of features, modelling techniques and computational frameworks are discussed and compared.

Chapter Three describes the effects of AMR, AMR-WB and AMR-WB+ speech compression on AER. The methodology describes compression methods, calculation of acoustic speech parameters and general AER framework. The modelling and classification of speech emotions is achieved using a benchmark approach based on the GMM classifier and speech features including TEO, MFCCs and GP-T&GP-F parameters. Experimental results are presented and discussed.

Chapter Four investigates effects of band reduction on the AER. Investigated speech bands include: 50Hz-250Hz, 250Hz–1kHz, 1kHz–4kHz, 1kHz–8kHz, 2kHz–8kHz, 4kHz–8kHz and the full band 0–8kHz. The modelling and classification of speech emotions is achieved using a benchmark approach based on the GMM classifier and speech features including TEO, MFCCs and GP-T&GP-F parameters. The experimental results are presented and discussed.

Chapter Five describes the combined effects of standard speech-compression techniques and simulated hearing loss on SER. The effects of the AMR, AMR-WB and AMR-WB+ codecs are compared against hearing-loss simulation emotion recognition from uncompressed speech. The recognition methods include techniques based on three different types of acoustic speech parameters: TEO features, MFCCs and GP-T&GP-F and GMM classifier. The experimental results are presented and discussed.

Chapter Six describes the effect of standard speech compression under noise conditions on automatic SER. Effects of Gaussian white noise addition to speech signals at SNR=15dB, 10dN and 5dB on SER is investigated. The recognition methods include techniques based on three different types of acoustic speech parameters: TEO features, MFCCs and GP-T&GP-F and GMM classifier. The experimental results are presented and discussed.

Chapter Seven investigates the effects of standard AMR, AMR-WB, MP3 audio and AMR-WB+ speech codecs on AER based on speech spectrogram features. The emotion recognition process is based not on speech acoustic parameters but on three types of speech spectrogram parameters: SS parameters, SS-CB and SS-Bark scale speech parameters. The experimental results are presented and discussed.

Chapter Eight investigates application of artificial bandwidth extension (ABE) to compressed narrow-band speech to test whether SER can be improved. The ABE is based on spectral folding and spectral envelope estimation. Modelling and classification of speech is performed with a benchmark approach based on the GMM classifier and a set of speech acoustic parameters including TEO, MFCCs and glottal parameters. The experimental results are presented and discussed.

Chapter Nine provides a summary and discussion of research findings and suggestions for future work.

Chapter 2: Literature Review of AER

2.1 Preview

This chapter provides an overview of previous studies related to AER from speech signals. AER [48] combines different research disciplines to generate efficient methodology that aims to automatically identify the affective state of speakers. It is the focus point of research aiming to improve human-computer interactions and to increase social acceptance of machine learning algorithms [47]. Examples of a large range of AER applications include for example mental state recognition, detection of deception, assessment of speaker's certainty, compassion and sincerity [41],[42],[48].

As indicated in numerous studies [43], [49], [32], [56], [39] emotions present in speech can have a significant effect on the accuracy of speech classification aiming either to recognise words (Automatic Speech Recognition) or identify a speaker (Automatic Speaker Recognition). This indicates that the knowledge of a speaker's emotional state could lead to an improvement of classification accuracy in both systems.

The majority of emotion recognition studies have focused on clean, uncompressed speech. Noisy environments and speech compression techniques used in modern communication systems have been shown to have a significant effect on acoustic speech characteristics [1], [2], as well as the accuracy of automatic speech and speaker recognition [32], [3], [4]. However, the effects of noise and speech compression on AER accuracy rates have not yet been sufficiently addressed.

This chapter starts with an overview of the existing body of research in AER. Mainstream methodologies are identified and explained. State of the art techniques are presented. The AER approach used in this thesis to investigate effects of noise and band reduction on AER accuracy is presented in the context of this review. In the second part of this review, a review of existing studies investigating the effects of noise and band reduction on AER accuracy are discussed.

2.2 Automatic Emotion Recognition (AER) from Clean Full Band Speech

The majority of existing AER techniques have been developed and validated using research databases of emotional speech. These databases contained recordings of full band (of at least 8 kHz) uncompressed speech, usually without or with very little noise. Research efforts demonstrated beyond doubt that speech signals can be used to efficiently determine the emotional states of speakers. Powerful numerical approaches have been proposed allowing the detection of speakers' emotions in a fully automatic way. The following paragraphs provide a brief review of AER methodology leading to the current state of the art methods.

2.2.1 AER from images and video sequences

AER systems have been designed to use a variety of single- or combined-mode signals including speech recordings, facial images, video recordings, audio-visual recordings as well as pulse, electrocardiogram (ECG) and electroencephalogram (EEG) signals [99]. For example, video sequences and still images showing human faces lead to particularly successful automatic emotion recognition. The current state of the art in this methodology belongs to recently published Microsoft Emotion API[123] . The system uses some form of Deep Learning [81], [113]. It takes a still facial image or video sequence of facial images as an input, and returns the confidence across a set of emotions. The method detects simultaneously anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. The system is open to online testing and development of user defined software applications. Although the classification accuracy varies from system to system and between databases, in general multi-modal systems improve the classification accuracy by 5% to 10% [100].

2.2.2 Advantages of AER from speech signals

When looking at reasons why using speech signals rather than other modalities such as facial images, ECG or EEG, we can say that speech is particularly good for the task of recognizing emotions due to the following reasons [49], [47].

-It is easily available and low cost signal. It can be captured in a discrete non-invasive way.

-Speech is arguably the most natural way of expressing emotions. These expressions are conveyed through both semantic expressions and acoustic modulation.

-Existing signal processing methods enable relatively easy acquisition of speech signals in real time.

-For other modalities, such as for example ECG or EEG, the acquisition process can be costly, invasive, and requiring specialised medical equipment.

2.2.3 Challenges of data collection for AER from speech signals

Automatic emotion recognition from speech signals brings a lot of challenges. The research shows many parallel lines of investigation using alternative representations and definitions of basic emotions, labels and databases [47], [48], [19]. Issues such as linguistic, cultural and ethnical differences between verbal expressions of emotions, as well as gender and age related differences have not been yet fully investigated.

Before the AER can be conducted a representative database of speech signals with associated labels needs to be created. The accuracy of this data is of key importance to AER. One can say that AER is only as good as the training data. This is because the training set provides a "ground truth" against which the system accuracy is validated during the training process. As illustrated in Figure 2.1, the process of training data generation encounters a lot of ambiguities and challenges. Emotional speech recordings can contain emotions that are acted, artificially induced or completely natural [47]. Ultimately, the last type is the most desirable, however, it is also the most difficult to obtain for many practical and ethical reasons. Speech labels used in AER can be either categorical (e.g. happy, sad, angry, etc.) or given as valence and arousal values on continuous emotion and intensity scales, respectively [48], [47]. The emotion labelling procedures can be conducted either by self-reporting or by using external markers. In self-reporting, the labels are assigned during the speaking tasks or after the tasks are completed by listening to speech recordings. The advantage of using markers is that intra- and inter-speaker reliability can be assessed independently and data achieving low reliability scores can be discarded.

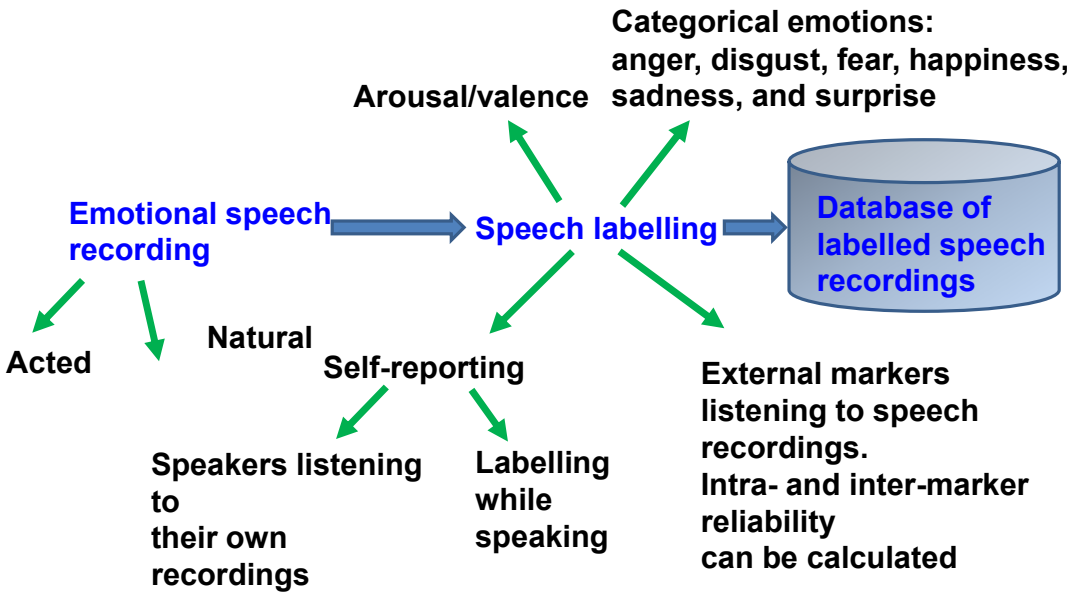

Figure 2.1 Training data collection and labelling.

2.2.4 Methodology of AER from speech signals

AER from speech signals is a classification problem. Therefore, like most of the classification tasks, it can be solved using standard pattern recognition techniques. These techniques generate emotional class models, which in turn can be used to classify (or label) a speech sample into a set of emotional categories or alternatively, placing it on a continuous scale of emotions [48].

A standard AER machine learning procedure is illustrated in Figure 2.2. It comprises a feature extraction procedure calculating low or high level acoustic speech parameters [48], [47]. This process is followed by a feature-selection (or data dimensionality reduction) procedure. Speech features are used in the training stage to generate emotional class models. These models are then applied in the testing stage to classify (or label) query speech samples into different emotional categories [48], [56].

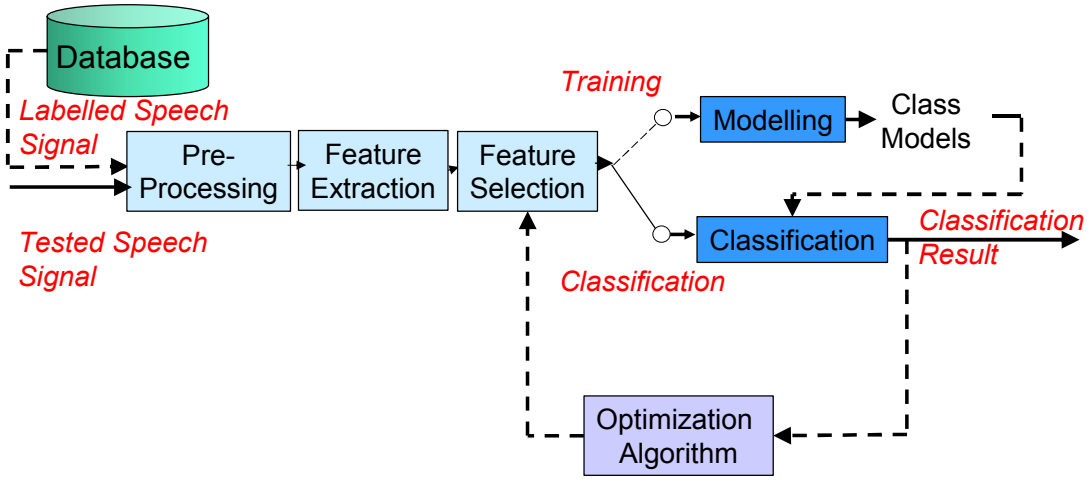

Figure 2.2: Standard machine learning approach used in AER.

Most often used features include low-level parameters such as fundamental frequency (F0), formant frequencies, time and frequency parameters of the glottal wave, and time and frequency parameters of the waveform. High level features such as for example mel frequency cepstral coefficients (MFCC) and Teager energy parameters are usually calculated as derivatives of the low level features. The most often used modelling/classification techniques include: Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), K-Nearest Neighbours (KNN) and various types of neural networks (NN) [19], [48].

2.2.5 Speech feature parameters for AER

At the low levels, acoustic speech feature parameters used in AER are derived from the basic Faint model of speech production [68], [68]. Low level speech parameters (Figure 2.3) are derived from voiced speech, i.e. speech produced by vocal folds vibration. Unvoiced speech components are usually discarded. These low-level descriptors include parameters such as [19], [118]:

-Fundamental frequency (F0) given as the fundamental frequency of the vocal folds vibration [68].

-Formant frequencies F1, F2, F3, ... given as the resonant frequencies of the vocal tract structure of bones and cavities [68].

-Time and frequency domain parameters of the glottal wave [19], [21], [22], [23].

-Spectral energy characteristics of both speech and glottal waves [19], [68]

High-level descriptors (HLDs) include parameters such as [19], [118]:

-Mel-Frequency Cepstral Coefficients (MFCCs) [19], [32].

-Parameters derived from the Teager Energy Operator (TEO) [19], [18].

An extensive overview of acoustic speech parameters at both low and high levels can be found in [19] and [68].

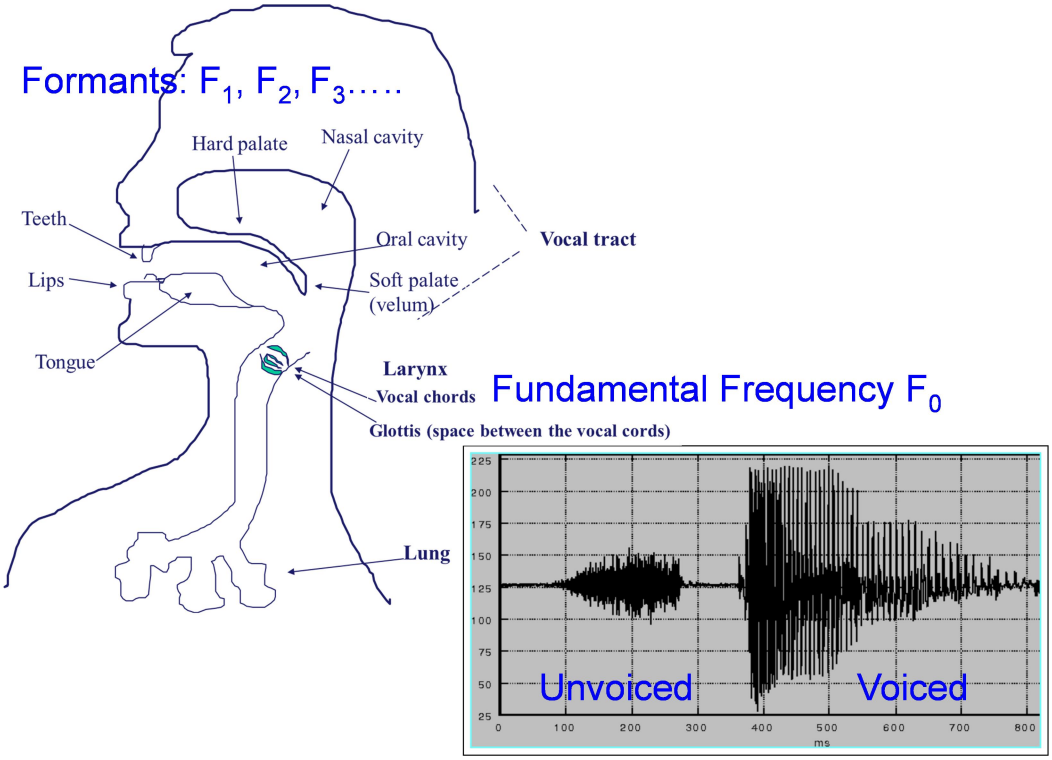

Figure 2.3: Low level speech parameters derived from voiced speech, i.e. speech produced by vocal folds . (Based on http: //z-pronunciation.weebly.com/speech-production.html accessed 10.05.2016.

2.2.6 Long and short term approaches to speech feature extraction for AER

Most of the reported work on AER from speech signals follows one of the following two standard feature extraction procedures [119], [48], [118]:

1. Long-time or turn based processing (also known as dynamic modelling), see Figure 2.4;

2. Short-time frame based processing, (also known as static modelling) see Figure 2.5;

Both methods extract speech parameters only from voiced speech. An extensive review of these two types of techniques can be found in [119].

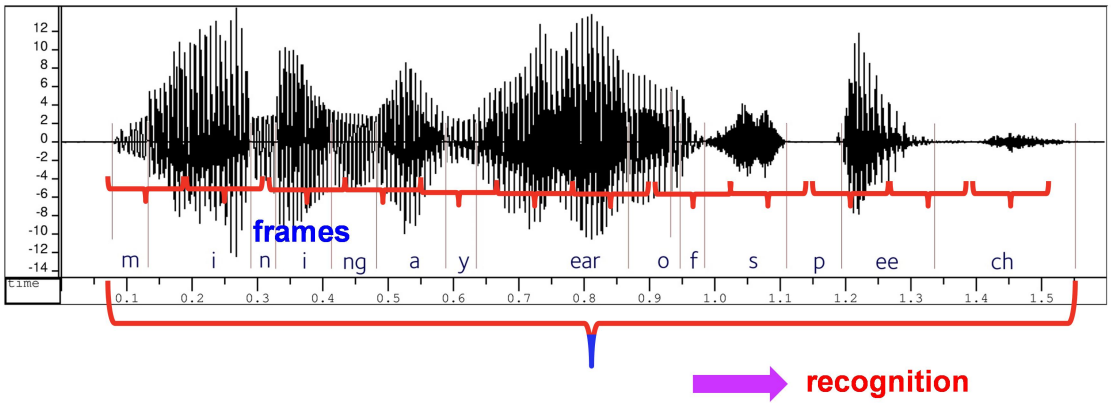

Figure 2.4 Long-time, turn based (static) processing of speech feature extraction.

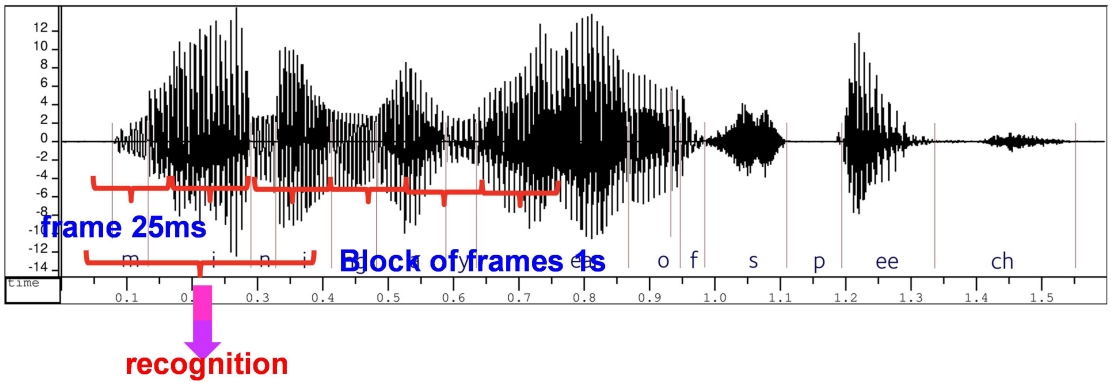

Figure 2.5 Short-time, frame based (dynamic) processing of speech feature extraction.

Over the last decade, the static turn-based approach has been used almost exclusively [119], [100], [101], although current trends are showing increased research interests in the frame-based approach [19], [21], [23], [24], [33]. In turn-based processing illustrated in Figure 2.4 emotions are recognized from concatenated voiced

components representing an entire sentence. The speech signal to be processed is first segmented into utterances using more or less complex methods usually based either on energy thresholds or estimation of signal tonality. Features are then extracted from short 20-30 milliseconds segments called frames and averaged over the whole sentence. This approach uses both low and high level acoustic parameters [119].

Given that the emotional states of speakers can change as frequently as at one second intervals [120], [121], the averaging used in static processing can easily remove vital intra-sentence information about emotional transitions. This disadvantage can be avoided by using a frame-based processing. This type of processing has been designed to recognize emotions by extracting features from very short 20-30 milliseconds frames and labelling emotional states occurring within short blocks of concatenated frames. These blocks typically include concatenated frames totalling in length anywhere between 0.5 seconds to 2 seconds [21], [20], [19], [23]. As illustrated in Figure 2.5, in frame-based processing a small number of frames are concatenated into a block and the classification is performed for each block separately. Unlike the turn-based method, the frame-based approach is capable of tracking fast intra-sentence changes of emotional states of speakers. Since recognition is performed every few seconds, the frame-based approach appears to be particularly suitable for applications when a rapid and ongoing assessment of emotional states is needed.

Over the last decade, the majority of research efforts in the area of AER have been devoted to selecting an optimal set of acoustic parameters [47], [48], [119], [116] allowing accurate classification. This is because the information of interest is actually encoded in the feature parameters and thus, the selection of appropriate features is the most important factor affecting the classification accuracy. The effect of classifier (or emotional class modelling technique) has been found to be of secondary importance, and most popular classifiers such as KNN, MFCC, SVM and HMM have been shown to provide very similar performance. An extensive comparison can be found in [19].

Many of the current approaches choose to use highly dimensional features sets combining large numbers of different feature parameters. It is believed that this approach can help cover all possible cues [117] allowing to differentiate between emotions. For example, a highly dimensional feature set described in [115] consists of over six thousand different parameters characterising speech acoustics. Classification

based on such highly dimensional data is computationally expensive, and time consuming. In addition, there is no guarantee that this approach is optimal. For example, as shown in [21], an application of a single energy parameter based on the Teager Enegy Operator significantly outperformed a number of highly dimensional feature sets in speech based detection of depression.

Calculation of feature parameters on a frame by frame basis can lead to extremely large number of data making the emotion recognition difficult to apply in many practical applications. Data compression or various optimal feature selection techniques such as for example Principal Component Analysis (PCA) or Minimum Redundancy Maximum Relevance (mRMR) [19] have been frequently applied in order to counteract this problem to some extent [102],[105]. These data reduction approaches not only reduce the computational complexity but also appear to remove data redundancy which in turn helps to achieve better classification results [21], [19].

2.2.7 AER approach used in the thesis

A classical machine learning approach (illustrated in Figure 2.2) based on acoustic speech parameters derived directly from low level features was applied consistently in all AER experiments described in this thesis. All experiments were conducted using the same Berlin Emotional Speech database [9]. This consistency allowed observing the effects of speech compression, band reduction and noise on AER within the same settings.

2.3 Automatic Emotion Recognition (AER) from Noisy and Limited Band Speech

At present, existing literature offers only very limited review of AER tested on noisy, narrow-band or other ways deteriorated speech signals recorded in real life conditions. To the best of our knowledge no systematic studies of effects of noise, band reduction or speech coding on AER have been conducted.

The following paragraphs review some of the more noticeable existing works related to this thesis.

2.3.1 AER from noisy speech

Most previous studies [82], [48], [83] in the spoken emotion recognition area focus on detecting emotional states in clean speech data easily recorded in quiet environment, but human beings are capable of perceiving emotions even in noisy environment. In recent years, robust emotion recognition in noisy speech has become an important issue in AER area since the real-world emotional speech signals are usually corrupted with different levels of noise.

Some efforts have been made for robust AER in noise by selecting a set of acoustic speech parameters optimised to perform well under noisy conditions. To mitigate negative effects of noise, Schuller et al. [84] proposed a fast noise-adaptive extraction of features sub-sets from a large set of 4000 acoustic features using an information gain ratio–based feature selection technique. In [85], [86], to reduce the influence of noise, a feature dimensionality reduction method called enhanced Lipschitz embedding was applied to map the extracted 64 acoustic features into a low-dimensional nonlinear manifold. Yeh and Chi extracted the joint spectro-temporal features from an auditory model and then applied them to detect the emotion status of noisy speech [87]. As far as the modelling and classification step of AER is concerned, very little has been done to improve resilience to noisy speech.

Over the last decade a new powerful theory of compressive sensing (CS) (also called compressive sampling) [88], [89], [90] has emerged. It was originally designed to address signal sensing and coding problems and has shown high potential in pattern recognition area. In particular, sparse representation in the CS theory has recently been used as a nonparametric classifier for pattern recognition and shows promising performance on face recognition [91], [92], [93] and speech recognition [94], [95] problems. The CS nonparametric classifier based on sparse representation is the so-called sparse representation classifier (SRC). In the SRC method, the test sample is represented as a sparse linear combination of the training samples, and the coding fidelity is measured by the l1-norm of coding residual. The sparse representation model of SRC assumes that, the coding residual follows the Gaussian distribution.

However, in reality, this assumption does not hold in a noisy environment. This is because the coding residual usually does not fit well into the Gaussian distribution. This implies that, the SRC may not be robust and effective in a noisy environment.

To improve the robustness and effectiveness of sparse data representation used in SRC, an enhanced (weighted) sparse representation modelling/classification (enhanced-SRC) method based on the maximum likelihood estimation has been proposed in [79] for robust AER in noisy conditions. It was tested to perform spoken emotion recognition, and its performance has been validated on both clean and noisy emotional speech. The effectiveness and robustness of the proposed method was investigated on clean and noisy emotional speech. The enhanced-SRC was compared with six typical classifiers, including linear discriminant classifier, K-nearest neighbour, C4.5 decision tree, radial basis function neural networks, support vector machines as well as sparse representation classifier [79]. Experimental results on two publicly available emotional speech databases, that is, the Berlin database and the Polish database, demonstrate the promising performance of the proposed method on the task of robust emotion recognition in noisy speech, outperforming the other used methods. In particular when looking at the results of simultaneous recognition of 7 different emotions reported on the Berlin Emotional Speech data [9] (also used in this thesis) at signal-to-noise ratio (SNR) values of 15dB, 10dB and 5dB (also used in this thesis), the following observations were made [79]:

-for clean speech (no noise) conditions, the SRC approach provided an average accuracy of 74.39% and the enhanced-SRC increased it to 81.68%;

-for noisy conditions with SNR=15dB, the SRC approach provided an average accuracy of 70.09% and the enhanced-SRC increased it to 80.13%;

-for noisy conditions with SNR=10dB, the SRC approach provided an average accuracy of 69.72% and the enhanced-SRC increased it to 79.35%;

-for noisy conditions with SNR=5dB, the SRC approach provided an average accuracy of 68.13% and the enhanced-SRC increased it to 77.74%;

This means that, the enhanced-SRC has helped to improve the AER accuracy by about 10% compared to the SRC model (classifier).

Further research of noise-resilient modelling techniques for AER lead to a slightly different approach extending upon the weighted sparse representation model (enhanced-SRC) introduced in [79]. This new method was presented in [80]. A new classification method for AER proposed in [80] used locality-constrained kernel

sparse representation-based classification (LC-KSRC) approach. The LC-KSRC was shown to be very efficient at learning discriminating sparse representation feature coefficients for AER, since it integrated both sparsity and data locality in the kernel feature space. Like in [79], the method proposed in [80] was compared with six representative emotion classification methods including linear discriminant classifier, K-nearest neighbour, radial basis function neural networks, support vector machines, sparse representation-based classification and kernel sparse representation-based classification. Experimental results were validated on two publicly available emotional speech databases, i.e., the Berlin database and the Polish database. When looking at the results of simultaneous recognition of 7 different emotions reported on the Berlin Emotional Speech data (also used in this thesis) for clean speech the SRC approach provided an average accuracy of 76.72% and the LC-KSRC increased it to 85.65%. Unfortunately, no LC-KSRC results were reported on noisy speech.

2.3.2 Effects of telephone band reduction and speech coding on speech signals

2.3.2.1 Effects of telephone band reduction on speech signals

The human voice has a frequency range from 20 Hz to 20K Hz. However, the traditional landline telephone bandwidth range has frequencies from 300 Hz to 3.4 K Hz which contains the most important frequencies for speech intelligibility, but a significant part of the spectral content of speech is lost [1], [2], [3]. In addition to bandwidth limitations of landlines, modern communication devices including mobile phones can also apply various coding techniques reducing the speech bandwidth or performing selective band reduction depending on subjective auditory perception criteria. These modifications can lead to changes in linguistic and emotional speech characteristics leading to significant deterioration of AER.

Due to the lack of systematic studies investigating effects of speech compression and band reduction on emotion recognition, there is no clear understanding of how AER would perform when applied to speech collected from existing speech communication devices. However, due to ubiquity of speech compression applied to modern communications, there is a need to investigate these issues and develop robust speech classification techniques that perform well not only in ideal uncompressed speech conditions, but also when using various types of speech codecs.

AER studies have been predominantly focused on uncompressed speech. Speech compression techniques used in modern communication systems have been shown to have a significant effect on acoustic speech characteristics [1], [2], as well as the accuracy of automatic speech and speaker recognition [3], [4]. It is therefore plausible that speech compression and related band reduction can have similar detrimental effects on AER. However, the effects of speech compression on AER rates have not been comprehensively addressed.

Kuenzel (2001) [98] investigated the effect of landline speech transmission on characteristics of speech signals was investigated. In particular, artefacts introduced by the removal of low-frequency range below 300Hz by a band-pass filter (300Hz–3kHz) from the transmission channel on vowel formants was investigated. The attenuation of speech frequency components with frequency decreasing towards 0Hz resulted in the slope of the transmission filter decaying towards low frequencies. Therefore the first speech formant F1 of most vowels was expected to be affected the most. Attenuation of the lower frequency formant F1 would result in increase of the relative weights of higher frequency formants. This would disturb the natural balance between formants and cause an artificial upward shift of its centre frequency of vowel formants. These theoretical predictions were tested experimentally using samples of speech transmitted through a telephone line. The samples included speech of 10 males and 10 females. Formant analysis of these speech samples has confirmed that the predicted effect on F1 has in fact occurred in all test samples. It applied to all vowels except /a/, whose F1 frequency was too high (equal or slightly above 300Hz) to be affected by the slope of the band-pass. It was concluded that the consequences of measurement errors arising from such artefacts could significantly reduce speech and dialect recognition as well as speaker identification.

In [1], the effect of mobile phone transmission involving speech bandwidth reduction to (200Hz–3.4kHz) on vowel formant frequencies was investigated. This investigation is interesting from the AER perspective since significant changes in formant frequencies could have an effect on emotion recognition from speech signals. Six male and six female speakers read a short passage into a mobile phone. Two simultaneous recordings were made, one at the far end of the phone line and the other via a microphone directly in front of the speaker. Measurements of formants F1, F2 and

F3 [68] were taken from between 15 and 25 stressed vowels per speaker in both sets of recordings. Due to the filtering effect of the phone transmission, F1 frequencies for most vowels were found to be higher than their counterparts in the direct recordings. The overall effect of the mobile phone on F1 frequencies was considerably greater than the landline telephone effect found by Kuenzel et al. [98]. Thus, on average the F1 values in the mobile condition were 29 per cent higher than in the direct condition. On the whole F2 measures were not significantly affected, in line with Kuenzel's [98] findings. F3 frequencies were also generally unaffected by the mobile phone transmission. Exceptions were found, however, particularly for individual speakers with relatively high F3s. In these cases the mobile recordings tended to yield significantly lower values.

Investigation conducted in [1] for voice communication over mobile phones was revisited in more recent study [96], [97]. Presented results showed an analysis of the long-term average speech spectrum (LTASS), the long-term formant distribution (LTF) in voiced sounds, and vowel formants F1, F2 and F3 of six speakers in five modes of mobile phone usage [96]. These modes were:

-normal holding of a mobile phone (NOR),

-with a bonbon (sweet) in the mouth (BON), with a cigarette between the lips (CIG),

-with the mobile phone between cheek and shoulder (SHO) and

-with the hand covering the mobile phone and mouth (HAN).

The results showed that each mode has an impact on spectral features and that the modes HAN and SHO have the greatest impact. The most striking results are the relative displacement of F1, which can reach 30% (e.g. vowel /a/ in HAN mode for males), formant F2, near 15% (vowel /i/ in SHO mode for males), and formant F3, about 5% (vowel /u/ in CIG mode for females). These findings suggested that forensic practitioners should exercise caution in interpreting formant measurements in speaker identification cases involving mobile phone transmission.

2.3.2.2 Effects of compression (coding) on speech signals

Similar effects on speech quality to those caused by band limitation of modern transmission systems can be expected from speech compression (coding) methods used to reduce the amount of bit transmission. This is because coding methods perform selective data compression procedures within different frequency bands based on various objectives and/or subjective criteria aiming to preserve basic acceptable level of speech intelligibility [44], [45]. These manipulations are likely to disturb natural relative values of formants and harmonic components of the fundamental frequency F0.

Speech compression has many industrial advantages for telecommunications and speech technology, which support and serve speech recognition for human-to-machine communications [44], [45]. These advantages include a reduction in the delay of data transmission using telephony, a reduction in the memory size needed to save speech recordings, and the memory of mobile phones. Thus, due to the ubiquity of speech compression applied to modern communications, there is a need to develop robust speech classification techniques that perform well not only in ideal uncompressed speech conditions, but also when using various types of speech codecs [38].

Effects of speech coding on formant information of speech signals has been researched in [3]. The Adaptive Multi-Rate (AMR) codec was standardized for the Global System Mobile Communication (GSM) network in 1999. It is also the mandatory speech codec to the Third Generation Wide Band Code Division Multiple Access (3G WCDMA) systems. Its use in digital cellular telephony is already widespread.

The study described in [3] reported results of examining the impact of the narrowband version of AMR codec, at its various bit rates, on acoustic parameters in the speech signal important for the task of forensic speaker identification (FSI). The analyzed acoustic parameters were the first three formant frequencies. It was shown that though the impact on these parameters as a function of bit rate can be quite significant, there was not a consistent trend. However, there were clear gender differences, likely caused by differences in pitch, with higher pitch female speech being affected significantly more by the codec than that of lower pitch male speech. In general formant frequencies were decreased by the codec, particularly in the case of high-

frequency formants. In the light of these findings, it was noted that caution is needed when analyzing speech that has been transmitted over the cell phone network utilizing AMR codec.

The use of speech coding systems in the telephone network raised the question of their impact on formant frequencies, fundamental frequency trajectories and other acoustics features used for speaker identification. One of the first studies looking at effects of speech coding on speaker recognition has been described in [4]. Results of a routine forensic investigation of three common speech coding systems (CELP, LPC and GSM) on the pitch and formant frequencies of speech extracted from several dialect regions of the TIMIT Speech Corpus have been presented. Speech fundamental frequency (F0) and formant frequencies (F1, F2, F3) extracted from time aligned, uncompressed and compressed (coded) speech samples are compared to establish the statistical distribution of error attributed to the coding system. From the analysis of both the spectrograms and distributions of formant deviation it was observed that formant trajectories were degraded under all three coding systems. This degradation was particularly clear at time intervals where there were rapid formant transitions. The smallest formant degradation occurred in speech compressed with the GSM, then CELP, and the LPC-compressed speech was most significantly affected. Pitch frequency tracking was also degraded under CELP and LPC, but not so under GSM. In subjective terms, the voice quality sounded quite harsh under CELP and LPC, and not as noticeably altered under GSM. Whether this substantially affects the accuracy of listener's speaker identification has not been established as separated listening tests would have to be conducted. Formant bandwidths for F1 were found to be relatively unaffected, whereas F2 and particularly F3 showed significant shifts in mean value and were broadened as effect of speech compression. Observed bandwidth increase of formants was attributed to the loss of spectral information. It was argued, that perceptually the loss of bandwidth is likely to affect a listener’s ability to clearly identify vowels and diphthongs characterising individual speakers. It was concluded that time-dependent fine-detailed characteristics of both the speech source (F0) and the transfer function of the vocal tract (formants) undergo significant degradation due to speech coding.

The magnitude of deviation in formant frequencies and the resulting percentage error in inter-formant distances showed that speaker identification tasks based on inter-formant distances are significantly affected by speech coding systems. Further investigation of effects of speech coding on speaker recognition reported in [3] investigated the influence of GSM speech coding on text independent speaker recognition performance. Three existing GSM speech coder standards were examined using TIMIT database. Firstly, it was found that all three GSM coding techniques degraded significantly the perceptual speech quality of speech and subsequently the outcomes of speaker recognition. These findings were consistent with [4]. Secondly, it was observed that a low LPC order in GSM coding was responsible for most performance degradations of speaker recognition.

2.4 Conclusion

This chapter presented a literature review of studies closely related to the topic of this thesis. In the first part of this chapter, a review of existing state-of-the art methodology for AER was presented. A comparison of various AER approaches and their results was given. A general pipeline framework of AER speech processing was described. This framework is consistently used throughout the thesis in AER experiments described in Chapters 3-8. In the second part of this chapter, a review of research works investigated effects of noise, telephone band reduction and speech coding on the quality of speech signals and speaker recognition/verification tasks. The described studies were consistent in the following general observations and conclusions.

Effects of noise on AER

-A decrease of SNR from 15dB to 5dB lead to decrease of AER [79];

-Detrimental effects of noise on AER can be to some extent reduced by either use of an optimal (noise resilient) sub-set of feature parameters [84] or by an application of a robust modelling/classification techniques [79].

Effects of telephone band reduction on speech characteristics

-Reduced telephone band (300Hz-3kHz) resulted in attenuation of the first formant F1 and to smaller degree formant F2 [1], [98];

-Higher formants were generally not significantly affected [1], [98];

-Attenuation of the lower frequency formant F1resulted in increase of the relative weights of higher frequency formants [1], [98].

Effects of speech compression (coding) on speech and speaker recognition

-Narrow band AMR coding lead to a significant impact on speech parameters as a function of bit rate, however there was not a consistent trend [3];

-Clear gender differences were observed due to AMR compression, likely caused by differences in pitch, with higher pitch female speech being affected significantly more by the codec than that of lower pitch male speech [3];

-In general formant frequencies were decreased by the AMR codec, particularly in the case of high-frequency formants [3];

-Coding systems (CELP, LPC and GSM) degraded significantly the perceptual speech quality of speech (formant and F0 trajectories) and subsequently the outcomes of speaker recognition [3], [4];

As noted in this review there is a clear lack of studies extensively examining effects of the above factors on AER. It is reasonable to expect that the reported changes in speech characteristics are likely to result in significant deterioration of AER from noisy and/or narrow band coded speech. The following chapters of this thesis are investigating this hypothesis.

Chapter 3: Effect of Speech Compression on AER

3.1 Preview

This chapter investigates the effects of standard speech compression techniques on the accuracy of AER. The effects of AMR, AMR-WB, AMR-WB+ and MP3 speech codecs are compared against emotion recognition from uncompressed speech. The recognition methods include techniques based on three different types of acoustic speech parameters: TEO-PWP features, MFCCs and GP-T&GP-F features. The results show that, in general, all four speech compression techniques led to a reduction in emotion recognition accuracy. However, the amount of degradation varied across compression methods and types of acoustic features.. The amount of degradation due to compression varied across compression methods, compression rates and genders. The accuracy of emotion recognition using the AMR-WB codec was higher than that of AMR, AMR-WB+ and MP3.

3.2 Introduction

Automatic recognition of emotions in speech has many applications in various human–machine communication systems, speaker recognition and verification, biometric security purposes, and medical and physiological services. The majority of emotion recognition studies have focused on uncompressed speech. Speech compression techniques used in communication systems have been shown to have a significant effect on acoustic speech characteristics [1], [2], as well as the accuracy of automatic speech and speaker recognition [3], [4]. However, the effects of speech compression on AER rates have not yet been addressed.

3.3 Advantages of Speech Compression Introduced to Industry

Speech compression has many industrial advantages for telecommunications and speech technology, which support and serve speech recognition for human-to-machine communications. These advantages include a reduction in the delay of data

transmission using telephony, a reduction in the memory size needed to save speech recordings, and the memory of mobile phones. Thus, due to the ubiquity of speech compression applied to modern communications, there is a need to develop robust speech classification techniques that perform well not only in ideal uncompressed speech conditions, but also when using various types of speech codecs.

3.4 Possible Factors Associated with Speech Compression

One possible factor associated with speech compression that could affect AER is spectral modifications to speech signals introduced during the coding and decoding procedures. Another important factor is the limited bandwidth used by some coding techniques. These factors can dramatically alter acoustic speech characteristics and directly affect the accuracy of emotion recognition in speech. In [4], the CELP and LP-based GSM speech codecs were shown to have a negative effect on the estimation of the fundamental frequency (F0). It was observed that the speech compression algorithm led to an increase in the F0 value of up to 30 Hz, making it closer to the F0 extracted from landline uncompressed speech [5]. Further, the F1 formants of vowels extracted from compressed speech were higher than those extracted from uncompressed speech [1]. In particular, [2] showed that F0 and formant frequencies (F1–F3) decreased significantly when estimated from speech compressed using the GSM AMR speech codec. Interestingly, not all acoustic speech features perform worse with compressed speech. For example, speech recognition accuracy has been shown to improve when using speech features such as the MFCC estimated from speech compressed using the GSM speech codec in comparison with uncompressed speech [3], [6]. Some of the limitations of these studies were the use of only the narrow-band GSM AMR speech codecs (300–4300 Hz) and a focus on only classical speech features in the analysis of the effects of speech compression. Despite the recent interest in AER research, no studies have comprehensively investigated the effects of speech compression on the affective characteristics of speech.

3.5 Chapter Aims

This chapter aims to address this research gap and investigates how standard speech compression techniques affect the accuracy of AER [38]. This study extends previous investigations into the effects of coding methods based not only on narrow-band

AMR, but also on AMR-WB and AMR-WB+. The effects of these codecs are analysed using a range of different features that have recently been reported to provide high performance in SER [7], [8], [38]. These features include TEO-PWP, MFCCs and GP-T&GP-F.

3.6 Method

3.6.1 Speech Database

The emotion recognition experiments were conducted on the Berlin Emotional Speech (BES) database described in [9]. The database contains speech samples that represent seven emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]) with linguistically neutral content. The sampling frequency of the speech samples was 8kHz. Table 3.1 presents the number of available speech samples for the different emotions.

Table 3.1: Description of the Berlin Emotional Database

	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Male	60	34	8	26	21	38	17
Female	67	45	30	29	37	40	36

3.6.2 Experimental Framework

The speech samples representing either compressed or uncompressed speech were normalised into the range ±1. After the removal of noise and detection of voiced/ silence, voiced speech frames were concatenated and used in the two-stage processing shown in Figure 3.1. In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.

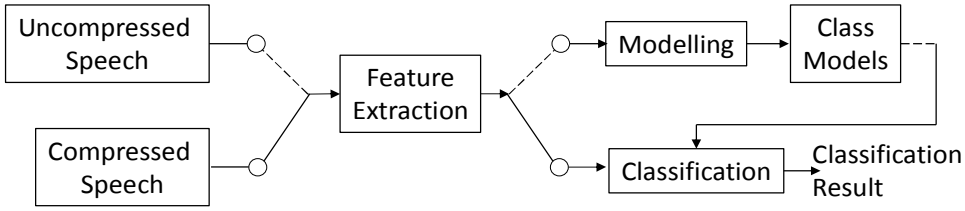

Figure 3.1: Experimental framework effect of compressed speech on SER

3.6.3 Average Percentage of Identification Accuracy

For both compressed and uncompressed speech, and for each feature/classifier combination, the training and classification process was run 15 times, each time with different training and testing sets selected using a stratified training and testing data-selection procedure [10]. For each run, 80% of the data were used in the training process and 20% were used in the testing. The classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10]:

$$\boldsymbol{APIA = \frac{1}{N_r}\frac{N_C}{N_T}100\%} \quad \textbf{(3.1)}$$

Where N_C is the number of test inputs correctly identified; N_T is the total number of test inputs; and N_r is the number of repeated tests. Emotion recognition was tested for each gender separately. Table 3.2 shows the compression bit rates tested in the experiments. Note that the compression rates corresponding to R1–R8 differ between the different types of codecs. This needs to be taken into account when evaluating the experimental results described in Section 3.6.2.

3.6.4 Speech Compression Methods

AMR-NB speech codec [11], [12]: AMR is based on ACELP and has eight narrow-band modes (ranging from 300 kHz to 3400 kHz). Each of the eight codec modes applies different bit rates: (AMR475) 4.75, (AMR515) 5.15 (AMR59) 5.9, (AMR67) 6.7, (AMR74) 7.4, (AMR795) 7.95, (AMR102) 10.2 and (AMR122) 12.2 kbps. The speech is coded frame by frame with a frame size of 20 ms (160 speech samples at 8 kHz sampling rate). For each speech frame, the speech signal is analysed using an LP of order 10 to calculate the LP coefficients, adaptive codebook, fixed codebook parameters and the gains. Each frame is divided into sub-frames, and the mode

switches between subsequent sub-frames. The resulting multi-mode (multi bit rate) coding has been efficiently applied in many mobile applications and wireless networks.

AMR-WB codec [13]: AMR-WB is an extension of AMR with a wideband range of 50Hz–7 kHz and a sampling frequency of 16 kHz operating at nine bit rates: 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 and 23.85 kbps. Like AMR, AMR-WB is based on the ACELP coding technique. However, AMR-WB uses a sixteenth-order LP short-term prediction filter and, for each frame, the LP parameters, adaptive and fixed codebooks and the gains are calculated. These parameters are encoded and transmitted as the speech frame is divided into sub-frames. The adaptive and fixed codebook parameters are transmitted for every sub-frame.

AMR-WB+ codec [14]: AMR-WB+ extends the AMR-WB method by adding TCX, bandwidth extension and stereo. While AMR and AMR-WB are optimised for speech compression, AMR-WB+ is designed to work with both speech and audio signals. The AMR-WB+ audio codec processes input frames of length 2048 samples at internal sampling frequencies ranging from 12,800 Hz to 38,400 Hz. There are two basic sets of rates: one for mono and one for stereo recordings. The basic mono rates are: AMR-WB+ 208 bit/frame (10.4 kbps), AMR-WB+ 240 bit/frame (12.0 kbps), AMR-WB+ 272 bit/frame (13.6 kbps), AMR-WB+ 304 bit/frame (15.2 kbps), AMR-WB+ 336 bit/frame (16.8 kbps), AMR-WB+ 384 bit/frame (19.2 kbps), AMR-WB+ 416 bit/frame (20.8 kbps) and AMR-WB+ 480 bit/frame (24 kbps). This study only applied the mono rates because the BES database only contained mono recordings [9].

MP3 codec [52]: The MP3 is an audio system for digital audio format which uses a lossy data compression format. The mp3 is a common format for streaming, storage and standard of digital audio compression or transfer to playback the music on the most audio digital players. It has been primarily designed to reduce the memory size used for data storage by the factor of ten or more compared with the original size, with keeping a sound quality is nearly equal to uncompressed audio quality for almost listeners. The MP3 file can be created at higher or lower bit rates, with including higher or lower resulting quality. The compression works by reducing accuracy of certain parts of sound that are considered to be beyond the auditory resolution ability of most people. This method is commonly referred to as perceptual coding. It uses

psychoacoustic models and the mel scale to discard or reduce precision of components less audible to human hearing, and then records the remaining information in an efficient manner.

Table 3.2: Bit rates used in the experiments for different speech compression systems

Bit rates	AMR	AMR-WB	AMR-WB+	MP3
R1	4.75	6.6	10.4	8
R2	5.15	8.85	12	16
R3	5.9	12.65	13.6	24
R4	6.7	14.25	15.2	32
R5	7.4	15.85	16.2	40
R6	7.95	18.25	19.2	48
R7	10.2	19.85	20.8	56
R8	12.2	23.85	24	64
R9				80
R10				96
R11				112
R12				128
R13				156
R14				192
R15				224
R16				265
R17				320

3.6.5 Speech Features

The acoustic speech parameters were calculated on a frame-by-frame basis with a frame length of 256 samples and 50% overlap between frames. The following subsections explain the feature extraction techniques applied to both compressed and uncompressed speech.

3.6.5.1 Mel-Frequency Cepstral Coefficients (MFCCs)

The MFCCs are some of the most frequently used features because they are shown to have good performance in speaker recognition and emotion classification in

speech [82], [19], [20]. For each frame, the Fourier transform and the energy spectrum were estimated and mapped onto the mel-frequency scale. The DCT of the mel-log energies were estimated, and the first 12 DCT coefficients provided the MFCC values used in the modelling and classification process.

3.6.5.2 Teager Energy Operator Features (TEO-PWP)

Features derived from the TEO [26] have been previously applied in emotion [21], stress [20], [21] and depression [22], [23], [24], [25] classification systems. The process of calculating the TEO parameters followed the frame-based method introduced in [24], which calculates the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through the PWP analysis as close estimates of the critical bands characterising the human auditory system [27]. For each frame of length 256 samples, values of the TEO instantaneous energy of a given signal $x[n]$ were calculated using (2), as proposed by Kaiser [26]:

$$\boldsymbol{\Psi(x[n]) = x^2[n] - x[n+1]x[n-1]} \quad \textbf{(3.2)}$$

The instantaneous energy was then used to evaluate the TEO autocorrelation function values using (3) [19]:

$$\boldsymbol{R_{\Psi(x)}[k] = \frac{1}{2M+1}\sum_{n=-M}^{M}\Psi(x[n])\Psi(x[n+k])} \quad \textbf{(3.3)}$$

Where M is the number of samples in the given frame. After smoothing with cubic splines, the area under the autocorrelation contour was calculated for each frame within each of the 17 frequency bands.

3.6.5.3 Glottal Time and Frequency Domain Features (GP-T&GP-F)

Glottal features have been shown to provide efficient classification of emotion [26] and depression [22], [23], [24] in speech. An iterative adaptive inverse filtering algorithm (IAIF) based on discrete all-pole (DAP) modelling was used to generate the glottal wave, and the glottal parameters were calculated using procedures included in the TTK Aparat Toolbox [28]. GP-T was represented by parameters describing amplitudes, timing, and duration of the opening and closing phases of the vocal folds. GP-F included three different parameters calculated from the spectrum of the glottal wave [21], [23]. These parameters described the differences between the amplitudes of

the first and second harmonic components of the glottal wave, the ratio of the sum of amplitudes of the higher harmonics to the amplitude of the first harmonic, and the spectral decay of the glottal waveform.

3.6.6 Modelling and Classification Methods

The modelling and classification tasks were completed using the GMM algorithm, which has been effectively used in speech modelling in various speech recognition tasks [28], [29], [30], [31], [9]. A GMM of order *M* models the probability density function of data as a weighted sum (or mixture) of *M* different Gaussian densities. Each Gaussian density has its own weight, mean and covariance. The EM algorithm was applied to estimate the optimal values of these parameters. The GMM (or training) stage was integrated with the Bayesian classification decision procedure, which determined the most probable classes for given query samples. A third-order GMM combined with the EM algorithm and the Bayesian classifier from the HTK toolbox were implemented to test the automatic classification of seven different emotional categories using compressed and uncompressed speech and different types of feature parameters.

3.7 Results and Discussion

This section describes how the three different speech compression techniques (AMR, AMR-WB, AMR-WB+ and MP3) affect the average multi-class emotion recognition accuracy performed when using three different types of features (MFCC, TEO-PWP and GP-T&GP-F). The results are presented in Figures 3.2 for male and female speakers.

3.7.1 Classification Outcomes for Uncompressed Speech

The emotion classification task, which aimed to distinguish simultaneously between seven different emotional classes, presented a significant challenge. The aim was to achieve results that did not fall below the pure guess level, which in this case was around 15%. The classification results for uncompressed speech (see Figures 3.2) showed that there were generally no significant differences between genders in emotion classification based on the non-glottal parameters and the glottal time domain parameters. The MFCC parameters were around 73% (see Figure 3.2), TEO-PWP was

78% (see Figure 3.3) and GP-T was 55% (see Figure 3.2) of the classification accuracy. Although TEO-PWP showed the best performance, a good performance was also given by the MFCCs. GP-F outperformed GP-T in both genders (see Figure 3.2). GP-F features were significantly more effective with male voices than with female voices. In particular, GP-F had 75% accuracy for male voices (see Figure 3.2) and only 59% accuracy for female voices. These results were consistent with previously reported emotion recognition outcomes based on uncompressed speech [7], [32], [17], [33].

3.7.2 Effect of Narrow-Band Adaptive Multi-Rate AMR Compression on Emotion Classification

For the MFCCs, the AMR compression led to low classification accuracy of 40%–51% (depending on the compression rate) compared to uncompressed speech. There was a clear decrease in classification accuracy from 50% to 40%, with the bit rates decreasing from R8 (12.2 kbps) to R1 (4.75 kbps). An outstanding 51% accuracy was observed for R5 (7.4 kbps) in the case of male speech (see Figure 3.2). Generally, there were no significant differences in these trends across genders.

For TEO-PWP features, classification accuracy dropped to around 50% compared to uncompressed speech (see Figure 3.2); however, it was almost the same (flat) for all bit rates from R8 to R1. Similar to AMR, an outstanding 57% accuracy was observed for R5 (7.4 kbps) in the case of male speech (see Figure 3.2). No other significant differences between genders were observed.

For GP-T&GP-F, the frequency parameters of GP-F outperformed the time domain parameters GP-T in both genders, and the male voice classification achieved higher results than the female voice classification (see Figure 3.2). Interestingly, the lowest bit rate R1 (4.75 kbps) led to the highest performance (51% GP-T, 60% GP-F for males and 45% GP-T, 52% GP-F for females) for compressed speech. An increase of the bit rate from R2 to R8 showed lower, but almost flat, performance compared to R1. These trends were similar for both genders. For all three types of features, the AMR codec provided higher accuracy of emotion recognition for male voices than for female voices.

3.7.3 Effect of AMR-WB Compression on Emotion Classification

In the case of MFCC features, AMR-WB compression led to low classification accuracy of about 60% compared to uncompressed speech, and it remained at this accuracy level for all bit rates decreasing from R8 (12.2 kbps) to R1 (4.75 kbps). There were no significant differences in these trends across genders.

For TEO-PWP features, the classification accuracy slightly increased for the lowest bit rate R1 (6.6 kbps) to around 79% compared to uncompressed speech (see Figure 3.2). An increase in the bit rate from R2 (8.85 kbit/s) to R8 (23.85 kbit/s) showed a decreasing slope in classification accuracy from around 74% (R2) to 67% (R8). There were no significant differences in these trends across genders.

The TEO-PWP results appear to contradict the informal belief that the lower the compressed speech bit rates, the higher the speech degradation and hence the lower the accuracy of emotion recognition. However, it is important to remember that the speech coding techniques used in this study were optimised for maximum speech intelligibility rather than for preserving emotional content. Moreover, previous studies of depression and emotion classification based on uncompressed speech have shown that the performance of the TEO features is highly dependent on the signal bandwidth [34], [35], and that optimal feature selection, which is effectively a speech compression process, can lead to a significant improvement in emotion classification results [35], [36]. A similar improvement over uncompressed speech was reported in [3] and [6], where speech recognition accuracy was improved when using the MFCC coefficients estimated from speech compressed by the GSM codec. The current results show that the combination of the wideband condition associated with the AMR-WB 6.6 kbps compression and the TEO-PWP features is likely to provide an optimal configuration for highly accurate emotion recognition in speech.

The glottal features (GP-T&GP-F) in Figure 3.2 show a different performance for male and female speakers. For male speakers, GP-T slightly outperformed GP-F, with a consistent performance across all rates R1–R8, leading to 55% accuracy for GP-T and 52% for GP-F. In contrast, for female speakers, GP-F outperformed GP-T, with almost flat performance across all rates R1–R8, leading to 40% accuracy for GP-T and 48% for GP-F.

3.7.4 Effect of AMR-WB+ Compression on Emotion Classification

For the MFCC parameters, the AMR-WB+ codec showed performance trends similar to AMR-WB (see Figure 3.2). In all cases, classification accuracy was slightly higher than that of AMR, but lower than that of AMR-WB.

For TEO-PWP and AMR-WB+ in mono mode, performance was slightly higher, but in all trends similar to the AMR (see Figure 3.2), and with classification accuracy slightly increasing with the increasing bit rate from R1 (10.4 kbit/s) to R8 (24 kbit/s).

The glottal features derived from the AMR-WB+ compression for both GP-T and GP-F exhibited a similar performance, with almost flat accuracy (about 45% on average) across all rates R1–R8 (see Figure 3.2). There were no significant differences between genders.

3.7.5 Effect of MP3 Compression on Emotion Classification

While there were no clear differences between genders, a strong dependency on the type of features could be observed.

In the case of MFCCs, the MP3 compression provided the best performance reducing the accuracy of SER only by about 5% compared to the uncompressed speech. The mp3 was followed by the AMR-WB which reduced the accuracy by 10% and the AMR-WB+ with accuracy reduction of about 15%. The worst performing was the AMR method with accuracy reduction of 20-25% compared to the uncompressed speech

In the case of TEO parameters, the MP3 was by far the worst performing method reducing the SER accuracy by 40% compared to the uncompressed speech. The best performance was achieved for the AMR-WB with about 10% reduction and AMR-WB+ with about 20% reduction. The AMR was close to mp3 reducing the performance by almost 30%.

In the case of glottal time and frequency domain parameters both, AMR-WB and AMR-WB+ showed high performance reducing the accuracy by only 10% compared to the uncompressed speech. This is consistent with previous reports [2]. The AMR and MP3 showed lower performance reducing the accuracy by up to 20%.

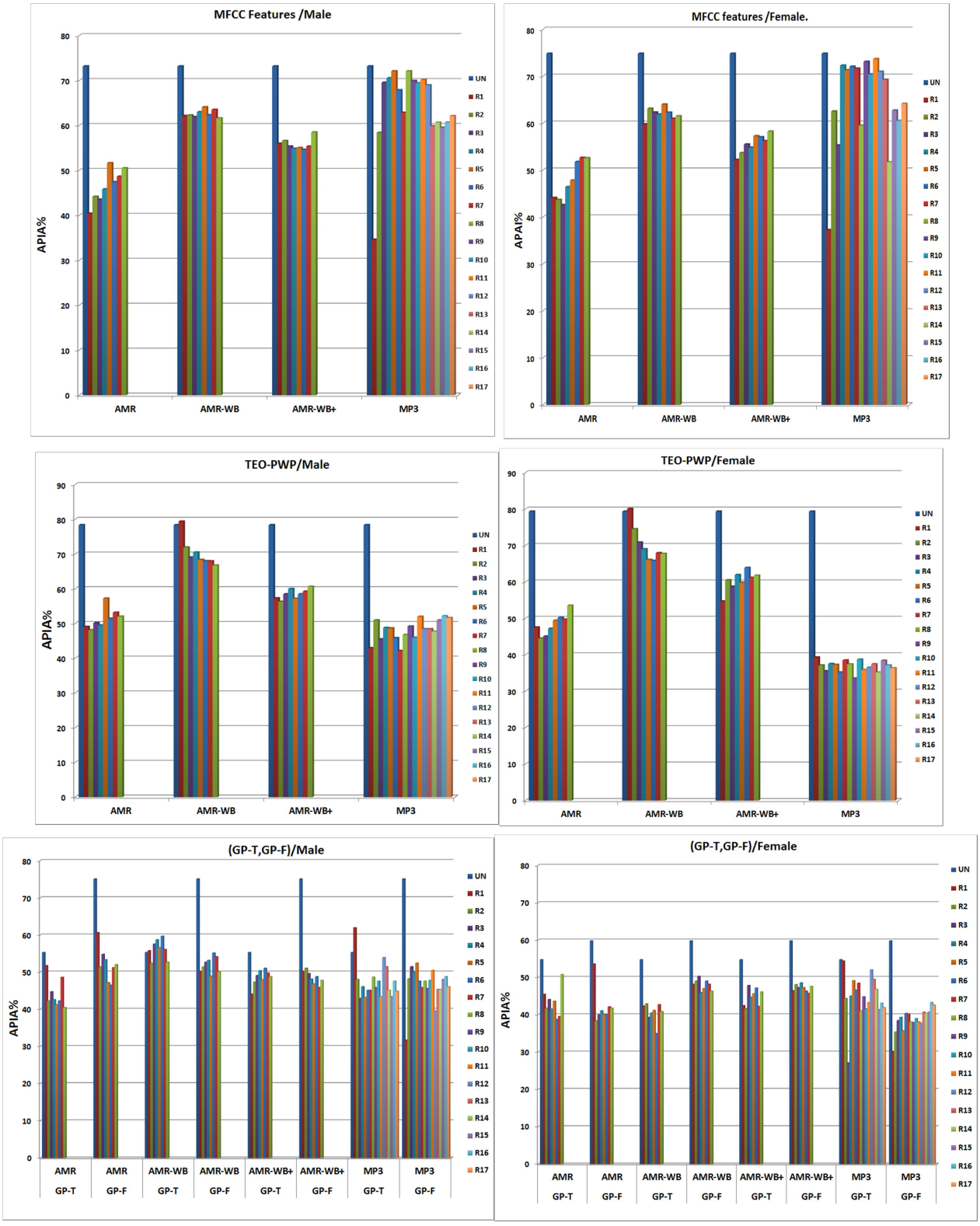

Figure 3.2: Average accuracy of multi-class emotion recognition for male and female speakers using MFCC, TEO-PWP and GP-T&GP-F features; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

3.8 Conclusion

This study investigated the effects of speech compression on the automatic simultaneous recognition of seven types of emotional speech samples obtained from the BES database.

The experiments included three different types of standard speech compression techniques (AMR, AMR-WB, AMR-WB+ and MP3) and three types of acoustic speech parameters (MFCC, TEO-PWP and GP-T&GP-F).

The modelling and classification of emotional speech was achieved using the GMM algorithm.

It is predictable that lower bit rates imply higher distortion to the speech signal; as such, they are expected to remove some information about speech emotions and lead to lower accuracy of AER.

In contrast, codecs with higher bit rates introduce less distortion and therefore could be expected to provide higher accuracy of AER.

AMR-WB was found to lead to smallest degradation of AER accuracy followed AMR-WB+ and MP3, the worse performance was observed for the narrow band AMR.

These observations indicated the importance of high frequency speech components to AER.

As expected, lower bit rates which imply higher distortion to the speech signal; lead to lower AER accuracy indicating that low bit rates remove vital emotional cues from speech signals. These cues are most likely to be located at high frequency end of speech spectrum.

The experimental results presented in this chapter confirmed this general expectation, showing that speech compression based on standard codecs degrades the AER outcomes.

However, the amount of degradation did not always increase with the decreasing bit rates.

In particular, the combination of the AMR-WB 6.6kbps compression and the TEO-PWP features provided an optimal configuration for high-accuracy multiclass emotion recognition in speech, thereby leading to results that were higher than for uncompressed speech.

The dependency patterns between the bit rates and emotion classification accuracy varied significantly across different genders, coding techniques and types of acoustic speech parameters used to distinguish between different emotions.

Generally, the classification results for all codecs and features, and across all bit rates, did not fall below 40%, which was significantly higher than the guessing threshold of 15% for the simultaneous recognition of seven classes of emotional speech.

One of the reasons for the observed degradation of emotional content in compressed speech could be the fact that current speech compression methods are optimised for maximum speech intelligibility.

Therefore, no objectives are used to ensure that the paralinguistic (emotional) content is preserved and fully conveyed to listeners. Future studies are needed to improve this aspect of speech coding standards.

There are two major questions arising from the presented here results.

(1) Why the MP3 compression works better with the MFCC features than with the TEO and glottal parameters?

(2) Why the AER based on the AMR-WB and AMR-WB+ outperforms AER based on the AMR and MP3?

The strong synergy between MFCCs and MP3 compression is most likely related to the use of the same perceptual mel frequency scale in both, derivation of the MFCC parameters and formulation of the MP3 compression criteria.

The mel scale [78] is a perceptual scale of pitches judged by listeners to be equal in distance from one another. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. When creating the MFCCs, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linear scale used to estimate TEO and glottal parameters.

This type of frequency warping allows for better representation of sound, and therefore it is used in the MP3 coding optimised for perceptually optimal audio compression.

High AER rates achieved with the MFCC indicate that the mel scale is not only preserving very well the melodic, but also the emotional cues of speech signals.

The differences between AER based on AMR, AMR-WB and AMR-WB+ can be attributed to the differences in signal bandwidth utilised by these techniques.

As shown in our experiments in Chapter 4, preserving the high frequency components (above 4kHz) of speech signals is essential for achieving high AER accuracy. Both AMR-WB and AMR-WB+ extend the AMR by including a wider range of high-end frequencies.

While the AMR-WB is optimized to preserve subjective qualities of speech signals, the AMR-WB+ is optimized for both speech and audio signals. This could explain why the AMR-WB outperformed slightly the AMR_WB+ in most cases of AER illustrated in Figure 3.2.

Finally, it can be observed that results described here are consistent with previously reported effects of speech compression on speaker recognition [3], [4].

In particular [3] reported that narrow band AMR coding lead to a significant impact on speech parameters as a function of bit rate, however there was not a consistent trend.

Clear gender differences were observed due to AMR compression, likely caused by differences in pitch, with higher pitch female speech being affected significantly more by the codec than that of lower pitch male speech.

Coding systems (CELP, LPC and GSM) degraded significantly the perceptual speech quality of speech (formant and F0 trajectories) and subsequently the outcomes of speaker recognition [4].

It is therefore likely that the observed changes in AER could be caused by similar mechanisms.

Chapter 4: Effect of Band Reduction on AER

4.1 Preview

This chapter investigates effects of band reduction on AER. As shown in Chapter 3, reduction of high frequency components of speech signals resulting from the narrow-band AMR compression led to high degradation of AER accuracy. Therefore, this chapter aims to determine contributions of different parts (sub-bands) of speech spectrum to the accuracy of AER.

The majority of emotion recognition studies conducted experiments on uncompressed full-band speech, where the term 'full band' refers to at least 8 kHz bandwidth.

In [7], [19], the effect of signal band energy contribution to stress and emotion identification was investigated to determine frequency ranges yielding the largest and smallest diversity between different stress levels and between different emotions.

Experiments based on speech data sampled at 8 kHz have shown that, in the case of different stress levels, the largest diversity between energy contributions occurred within low frequencies, ranging from 0Hz to 250Hz, and within high frequencies, ranging from 2.5kHz to 3.5kHz.

Similarly, in the case of different emotions, the largest diversity between energy contributions from different frequency bands occurred at the low-frequency range of 0Hz to 250Hz and the high-frequency range of 2.5kHz to 4kHz.

The middle range of frequencies (250Hz to 2.5kHz) did not show clear differences between stress levels or emotions.

In the case of stress classification, the 2.5kHz to 4kHz features yielded classification rates of 79%, whereas the whole bandwidth yielded 82%.

Similarly, in the case of emotion classification, the 2.5kHz to 4kHz features yielded classification rates of 79%, whereas the whole bandwidth yielded 87%.

The middle range of frequencies (250Hz to 2.5kHz) was much less efficient in the classification task, resulting in 51% correct classification rates for stress and only 37% for emotions.

The combined ranges of 0Hz to 250Hz and 2.5kHz to 4kHz were also found to be efficient, providing 80% correct classification rates for stress and 83% for emotions.

The experiments described in the following sections aim to validate findings described in [7] and [19] to confirm that high frequency components of speech signals play key role in AER.

4.2 Method

4.2.1 Speech Data

AER experiments were conducted on the BES database, which is one of the most frequently used standard AER testing databases [9]. The BES database contains speech samples representing seven categorical emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]). The text of each utterance was designed to be emotionally neutral, thus providing no linguistic cues about its emotional content. The aim was to recognise emotions using only acoustic cues. All files were available in wav audio format recorded at 16 kHz sampling rate and 16-bit amplitude resolution. Table 3.1 shows the number of available speech samples for the different emotions.

4.2.2 Experimental Framework

As shown in Figure 4.1, a standard signal classification pipeline has been applied. In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.In all experimental cases, the speech samples were first normalised into the range ±1.

After detection of voiced/silence, voiced speech frames were concatenated and used in the processing framework illustrated in Figure 4.1.

Before passing to the feature extraction stage, speech signals were convolved with a band limiting filter. After limiting the bandwidth, feature parameters were calculated. The experiments used three types of feature parameters: MFCCs, TEO-PWP parameters, as well as, glottal time (GP-T) and frequency (GP-F) parameters. The features were then applied to perform training leading to generation of acoustic class models representing anger, happiness, sadness, fear, disgust, boredom and neutral speech.

Acoustic class models were generated using the classical GMM modelling method with three mixtures [4]. The GMM software was obtained from the HTK toolbox [54].After generation of acoustic class models, the same process of band limiting and feature extraction was applied to test speech samples. The features were then passed to the Bayesian classifier typically associated with the GMM [68].

Classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10],In each case of gender, band limitation and type of features, a separate training and testing (classification) procedures were conducted [57].

In all experimental cases, the training and classification process was repeated 15 times, each time with different mutually exclusive training and testing sets selected using a stratified data-selection procedure [10]. For each repeat, 80% of data were used in training and 20% were used in testing.

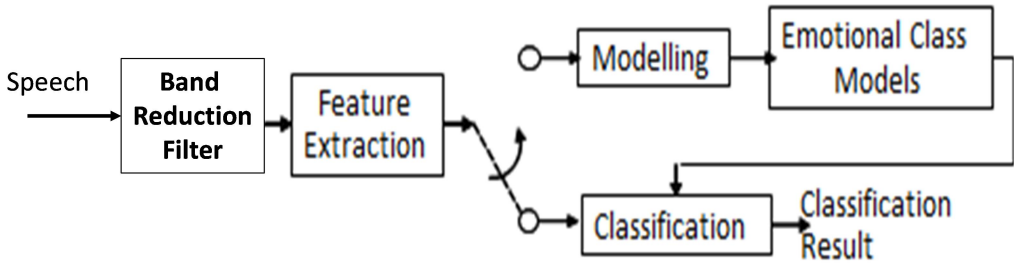

Figure 4.1: Experimental framework for testing effects of band reduction of speech on AER.

4.2.3 Bandwidth Reduction

The original speech bandwidth of 8 kHz was divided into seven sub-bands, B1–B7, as shown in Table 4.1, where B8 denotes the full band. Speech signals within each of these bands were tested separately to examine their individual contributions to the overall classification accuracy of speech emotions.

The sub-band speech was extracted by passing the speech signal through band-pass Butterworth filters designed using the DSP System Toolbox [50]. Features calculated within a given sub-band were then used in the emotional class modelling and speech classification procedures.

Table 4.1: Frequency bands used in emotion recognition tasks

B1	B2	B3	B4	B5	B6	B7
0.05–0.25k	0.25–1k	1–4k	1–8k	2–8k	4–8k	0–8k

4.2.4 Calculation of Speech Features for Emotion Recognition

Acoustic speech features used to differentiate between different emotions included MFCC [32], TEO-PWP parameters [19] and GP-T&GP-F domain parameters. All feature parameters were calculated on a frame basis with a frame length of 256 samples and 50% overlap between frames.

MFCCs: MFCCs have been reported to provide good performance in speaker recognition and emotion classification in speech [15], [16], [17]. For each frame, the Fourier transform and the energy spectrum were estimated and mapped onto the mel-frequency scale. The DCT of the mel-log energies was estimated, and the first 12 DCT coefficients provided the MFCC values used in the modelling and classification process.

TEO-PWP: Features derived from the TEO [18], [53], [7] have been previously applied in emotion, stress and depression [20], [22] classification systems. Calculation of the TEO parameters involves estimation of the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through PWP decomposition [53], [24] into close estimates of the critical bands. For each frame of length 256 samples, values of the TEO instantaneous energy of a given signal $x[n]$ were calculated using Equation 3.2 [26].

Instantaneous energy was then used to evaluate the TEO autocorrelation function values using Equation 3.3, where M was the number of samples in the given frame. After smoothing with cubic splines, the area under the autocorrelation contour was calculated for each frame within each of the 17 frequency bands.

GP-T&GP-F: Glottal features have been shown to provide efficient classification of emotion [33], [30] and depression [21], [22], [23] in speech. An IAIF algorithm based on DAP modelling was used to generate the glottal wave, and the glottal parameters were calculated using procedures included in the TKK Aparat Toolbox [27]. GP-T features were represented by parameters describing amplitudes, timing and duration of the opening and closing phases of the vocal folds. GP-F features included three different parameters calculated from the spectrum of the glottal wave. These parameters described the differences between amplitudes of the first and second harmonic components of the glottal wave, the ratio of the sum of amplitudes of the higher harmonics to the amplitude of the first harmonic, and the spectral decay of the glottal waveform.

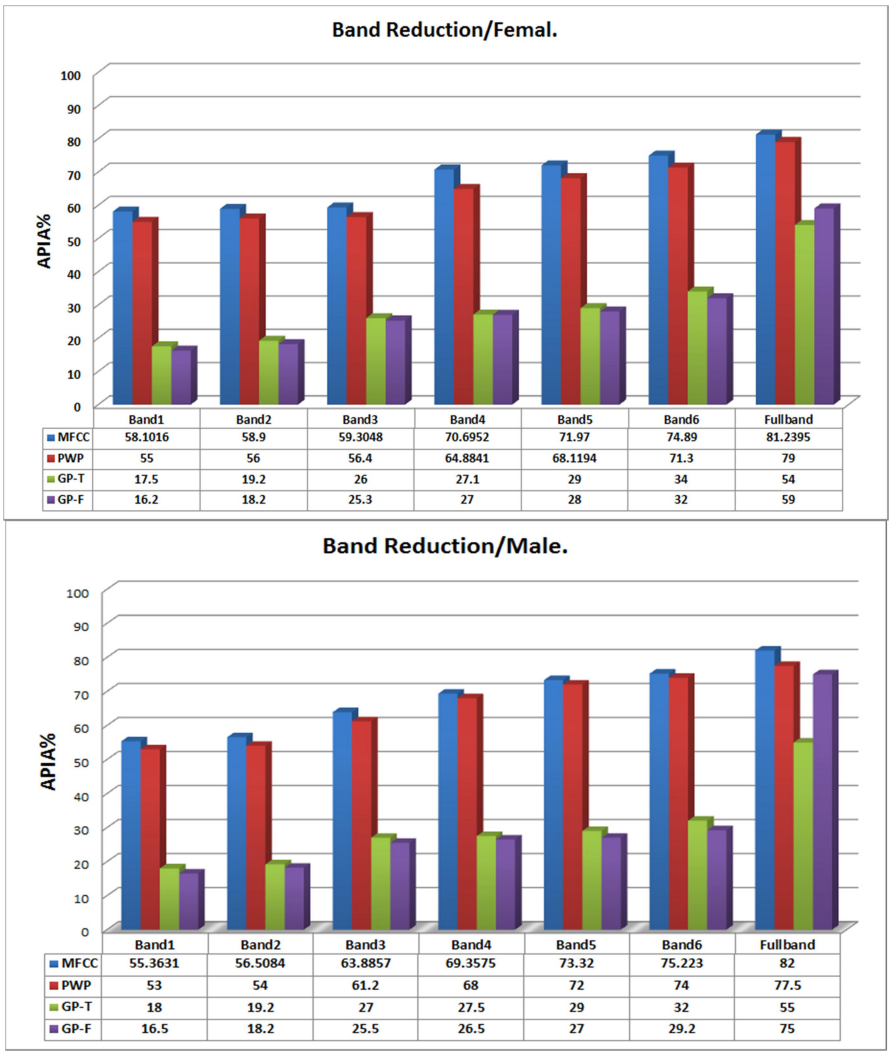

Figure 4.2: AER using limited-band speech.

4.3 Experiments and Results

The effects of band reduction for male and female speakers is shown in Figure 4.2.

For both genders, the MFCC features showed the highest performance across all bands, closely followed by the TEO parameters, which were only slightly worse. GP-T&GP-F showed significantly lower performance compared to MFCC and TEO parameters. These observations are consistent with [38].

Similarly, independent of gender, the full-band (0 kHz to 8 kHz) speech led to the highest AER accuracy.

However, the exclusion of the low-frequency range of 0 kHz to 1 kHz reduced accuracy by only 10%.

This could indicate that the low-frequency components containing the fundamental frequency information play an important role in AER.

When retaining only the high-frequency range of 4 kHz to 8 kHz, only a small reduction of around 10% in AER accuracy was observed.

This indicates that high-frequency components above 4 kHz also play an important role in AER. These observations are consistent with [38], [8].

In general, no significant gender-dependent differences were observed, apart from the fact that, for male speakers, the glottal frequency domain parameters with the full-band signal showed outstandingly high performance of around 72% accuracy, which was almost as high as the accuracy of the MFCCs (80%).

4.4 Conclusion

Possible factors associated with speech compression that could affect AER include spectral modifications to speech signals introduced during the coding and decoding procedures. Another important factor is the limited bandwidth used by some coding techniques. These factors can alter acoustic speech characteristics and directly affect the accuracy of emotion recognition in speech.

This chapter investigated how a band limitation affects the accuracy of AER. The effects of these factors on AER were analysed using a range of different features that have recently been reported to provide high performance in speech emotion recognition. These features include MFCCs, TEO-PWP and GP-T&GP-F parameters. Acoustic class models were trained and classified using the GMM technique.

The results indicated that the low-frequency components (0 kHz to 1 kHz) of speech containing the fundamental frequency information, as well as the high-frequency components (above 4 kHz) play an important role in AER.

Future investigations should include speech signals sampled at higher frequencies with bandwidth extending beyond 8 kHz. Effects of various auditory conditions limiting the

speech spectrum, such as hearing impairment and the use of hearing aids and cochlear implants on the AER, should be investigated.

Chapter 5: Effect of Speech Compression and Hearing Loss Simulation on AER

5.1 Preview

This experiment investigates the effects of standard speech compression techniques on AER from speech modified in a way that simulates a typical hearing loss. As indicated in Chapters 3 and 4, removal of high frequency components from speech spectrum leads to significant degradation of AER accuracy. Given that, a typical age-related hearing loss is characterised by reduced ability to hear high frequency components of speech, hearing aids users may not be able to capture full emotional contents of speech and subsequently experience reduced ability to recognise emotions from speech signals. This effect may be even larger when an impaired hearing person listens to compressed speech. Experiments conducted in this chapter indicated that this hypothesis could be true. However, this is only an indication based on machine learning. The full proof would require subjective listening tests which are beyond the scope of this thesis.

5.2 Hearing Loss

Hearing loss is seen by most Australians as an age related problem.

However, due to the noisy modern environment combined with the lifestyle choices involve listening to loud music, more Australians have a greater risk of acquiring a hearing loss earlier in life.

In 2005, an estimated 3.55 million Australians were believed to have hearing loss – that's roughly 17 percent of the total population [77]. In general, the amount of people with a permanent hearing loss increases significantly with age with males experiencing relatively higher levels of hearing loss.Hearing loss, also known as hearing impairment, is a partial or total inability to hear. A deaf person has little to no hearing [76].

Hearing loss may be caused by a number of factors, including: genetics, ageing, exposure to noise, some infections, birth complications, trauma to the ear, and certain medications or toxins. Hearing loss is categorized by type, severity. Furthermore, a hearing loss may exist in only one ear (unilateral) or in both ears (bilateral). Hearing loss can be temporary or permanent, sudden or progressive.

The severity of a hearing loss is ranked using hearing audiograms.The audiograms measure sound intensity above a nominal threshold that a sound must have before being detected by an individual.An audiogram is measured in decibels of hearing loss, or dB HL within octave frequency intervals covering a hearing range from 0 to 8kHz [76]. Hearing loss may be ranked as slight (16-25 dB HL), mild (26-40 dB HL), moderate (41-54 dB HL), moderately severe (55-70 dB HL), severe (71-90 dB HL) or profound (91 dB HL or greater) [76].

5.3 Method

5.3.1 Database

The emotion recognition experiments were conducted on the Berlin Emotional Speech (BES) database described in [9].

The database contains speech samples that represent seven emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]) with linguistically neutral content. The aim was to recognise emotions using only acoustic cues.

All files were available in wav audio format recorded at 16 kHz sampling rate and 16-bit amplitude resolution. Table 3.1 shows the number of available speech samples for different emotions.

5.3.2 Experimental Framework

As shown in Figure 5.1, a standard signal classification pipeline has been applied.

In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.

In all experimental cases, the speech samples were first normalised into the range ±1.

After detection of voiced/silence, voiced speech frames were concatenated and used in the processing framework illustrated in Figure 5.1.

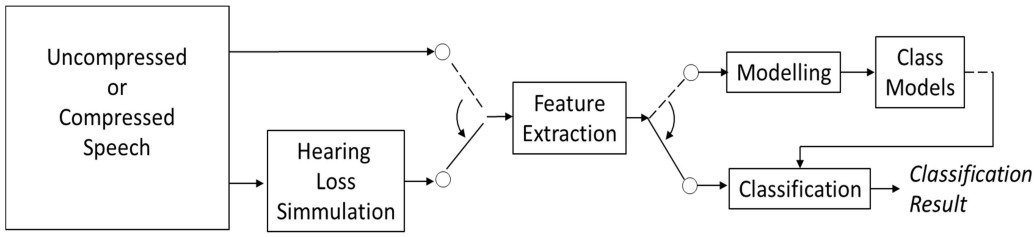

Figure 5.1: Experimental framework investigating effects of speech compression and hearing loss simulation on AER.

After applying a speech compression, the signal was passed through a filter simulating a typical age related hearing loss [76]. The speech was compressed using three different compression methods: AMR, AMR-WB and AMR-WB+ described in Chapter 3.The female and the male speech samples were tested separately to determine the effect of gender on the classification results. After passing the either compressed or uncompressed speech through a hearing loss simulator, feature parameters were calculated. The experiments used three types of feature parameters: MFCCs, TEO-PWP parameters, as well as, glottal time (GP-T) and frequency (GP-F) parameters.

The features were then applied to perform training leading to generation of acoustic class models representing anger, happiness, sadness, fear, disgust, boredom and neutral speech. Acoustic class models were generated using the classical GMM modelling method with three mixtures [4]. The GMM software was obtained from the HTK toolbox [54].

After generation of acoustic class models, the same process of hearing loss simulation and feature extraction was applied to test speech samples. The features were then passed to the Bayesian classifier typically associated with the GMM [68].

Classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10],

In each case of gender and compression method, a separate training and testing (classification) procedures were conducted [57].

In all experimental cases, the training and classification process was repeated 15 times, each time with different mutually exclusive training and testing sets selected using a stratified data-selection procedure [10].

For each repeat, 80% of data were used in training and 20% were used in testing.

5.3.3 Hearing Loss Simulation

The hearing loss filter was generated using the AngelSimTM(TigerCIS): Choclear Implant and Hearing Loss Simmulator [55], [75] software package. Which can be downloaded from http://www.angelsound.tigerspeech.com.

The frequency response of a low-pass filter simulating a typical age related high frequency hearing loss was setup using an audiogram interface illustrated in Figure 5.2.

The audiogram characteristics in Figure 5.2 show a slowly decaying loss of hearing towards high frequencies. The loss has been setup in a mild to moderate range.

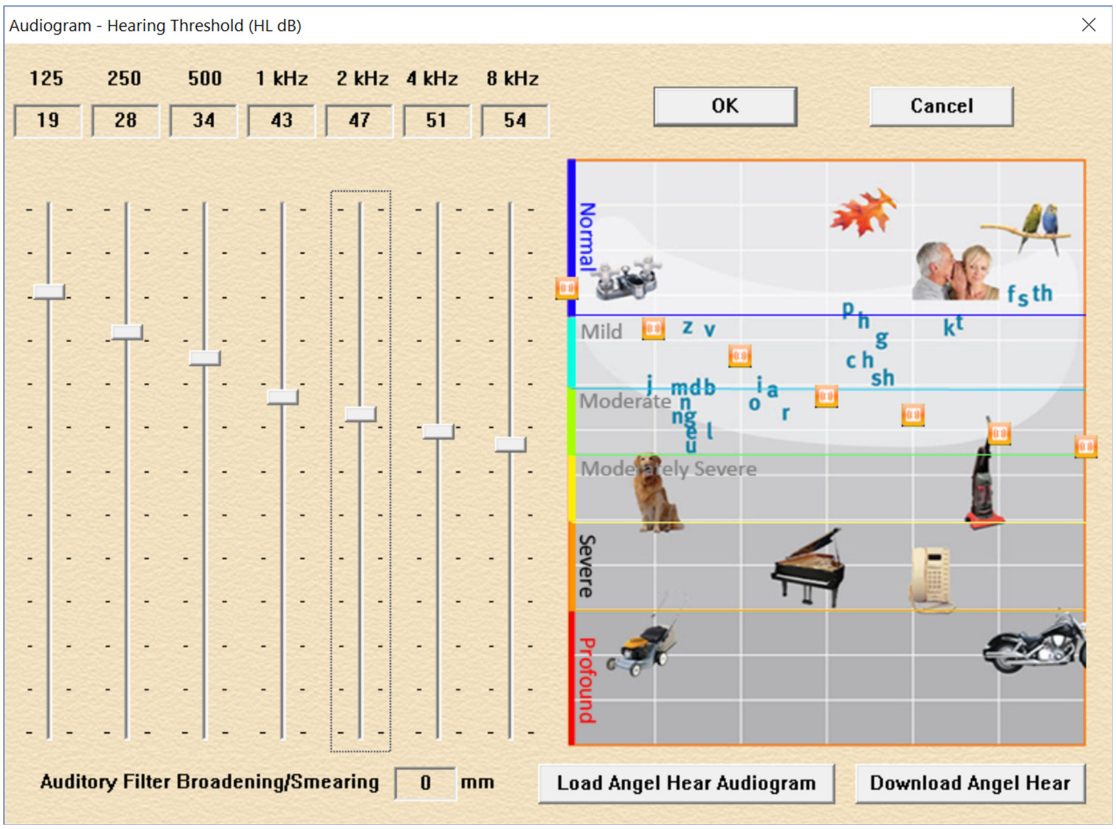

Figure 5.2: Audiogram setup used to generate a low-pass filter simulating a typical mild-to-moderate hearing loss. The audiogram and the corresponding low-pass filter were generated using publically available AngelSim software http://www.tigerspeech.com/angelsim/angelsim_about.html

5.3.4 Calculation of Speech Features for Emotion Recognition

Acoustic speech features, including MFCC [32], TEO-PWP [19] and glottal time GP-T and frequency domain GP-F parameters were used to differentiate between different emotions. All feature parameters were calculated on a frame-by-frame basis with a frame length of 256 samples and 50% overlap between frames.

MFCCs: MFCCs have been reported to provide good performance in speaker recognition and emotion classification in speech [15], [16], [17]. For each frame, the Fourier transform and the energy spectrum were estimated and mapped onto the mel-frequency scale. The DCT of the mel-log energies was estimated, and the first 12 DCT coefficients provided the MFCC values used in the modelling and classification process.

TEO-PWP: Features derived from the TEO [18], [53], [7] have been previously applied in emotion, stress and depression [20], [22] classification systems. Calculation of the TEO parameters involves estimation of the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through PWP decomposition [53], [24] into close estimates of the critical bands. For each frame of length 256 samples, values of the TEO instantaneous energy of a given signal x[n] were calculated using Equation 3.2 [26].

Instantaneous energy was then used to evaluate the TEO autocorrelation function values using Equation 3.3 [40]. where *M* was the number of samples in the given frame. After smoothing with cubic splines, the area under the autocorrelation contour was calculated for each frame within each of the 17 frequency bands.

GP-T&GP-F: Glottal features have been shown to provide efficient classification of emotion [33], [30] and depression [21], [22], [23] in speech. An IAIF algorithm based on DAP modelling was used to generate the glottal wave, and the glottal parameters were calculated using procedures included in the TKK Aparat Toolbox [27]. GP-T features were represented by nine different parameters describing amplitudes, timing and duration of the opening and closing phases of the vocal folds. GP-F features included three different parameters calculated from the spectrum of the glottal wave. These parameters described the differences between amplitudes of the first and second harmonic components of the glottal wave, the ratio of the sum of amplitudes of the higher harmonics to the amplitude of the first harmonic, and the spectral decay of the glottal waveform.

5.4 Results and Discussion

This section describes how each of the three speech compression methods (AMR, AMR-WB and AMR-WB+) affects AER accuracy from speech modified by a mild-to-moderate high frequency hearing loss filter.

The experimental results for both genders and for three compression methods are presented in Figure 5.3 (for AER using MFCC features), in Figure 5.4 (for AER using TEO-PWP features) and in Figure 5.4 (for AER using GP-T&GP-F).

The following sections describe and analyse the results separately for each of the three speech compression methods: AMR, AMR-WB and AMR-WB+.

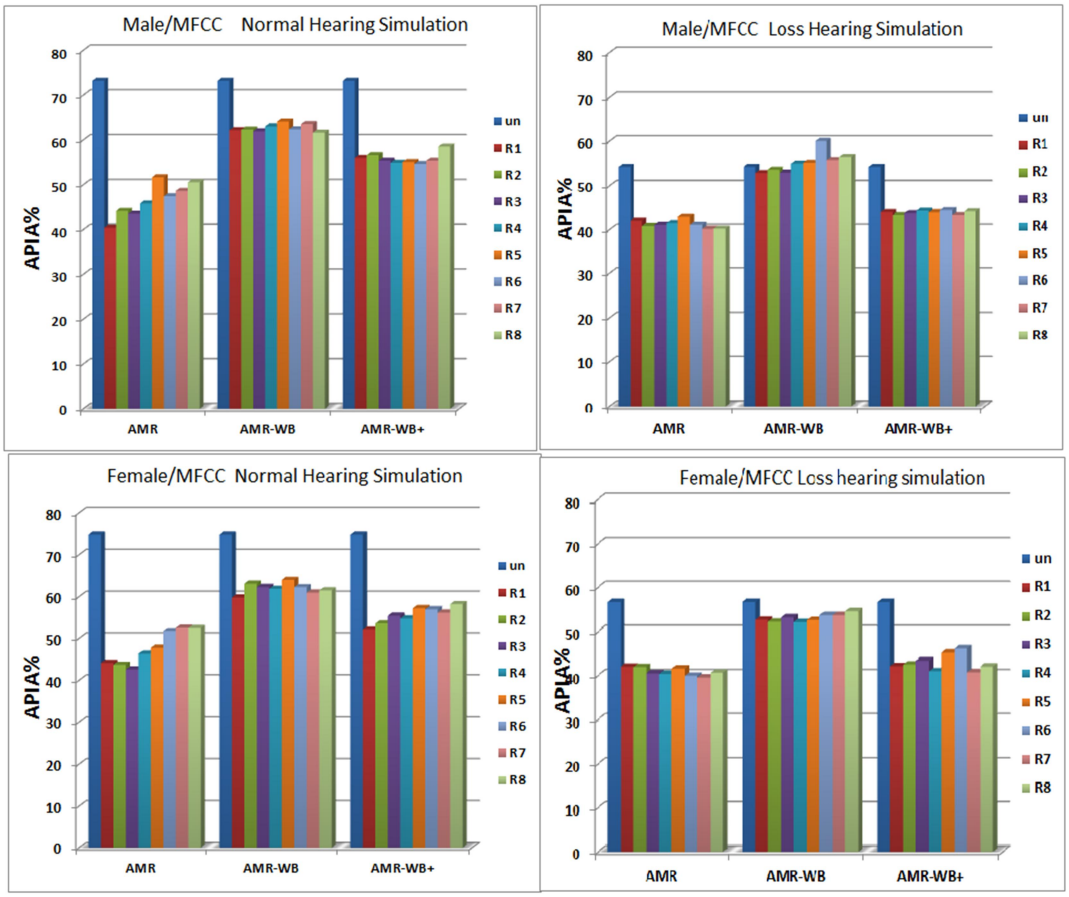

Figure 5.3: Average accuracy of multi-class emotion recognition for male and female speakers using MFCC features; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order for both normal and hearing-loss simulation

5.4.1 Effect of AMR and Hearing Loss Simulation on AER

MFCC (Figure 5.3)

Without hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech without HLS was about 72% for both genders.

For AMR without HLS, the AER accuracy from compressed speech increased from 40% to 50% with increasing compression rate for both genders.

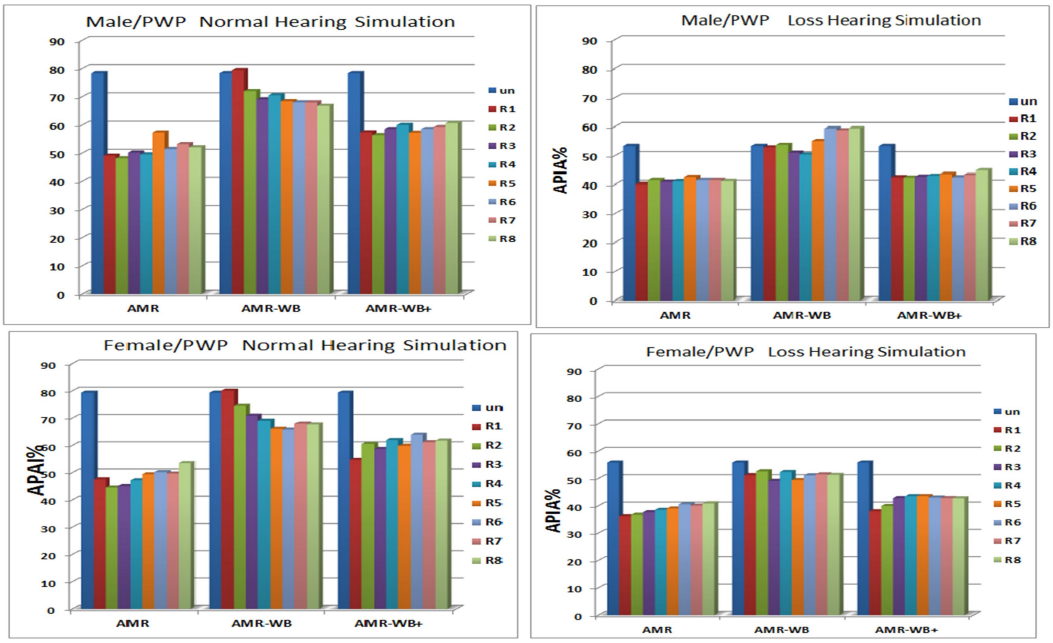

Figure 5.4: Average accuracy of multi-class emotion recognition for male and female speakers using TEO-PWP features; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order for both normal and hearing-loss simulation

With hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech with HLS was about 55% for both genders.

For AMR with HLS, the AER accuracy from compressed speech led to relatively low classification accuracy of 40% (almost flat across compression rates) for both genders.

It can be therefore concluded that when using MFCCs, the AMR compression reduced the AER accuracy by about 22%-32% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 10% and made it flat across compression rates.

TEO-PWP (Figure 5.4)

Without hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech without HLS was about 78% for both genders.

For AMR without HLS, the AER accuracy from compressed speech lead to 50% flat accuracy across all rates for males and an increasing accuracy from 45% to 50% with increasing compression rate for females.

With hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech with HLS was about 50% for males and 55% for females.

For AMR with HLS, the AER accuracy from compressed speech was about 40% for males and 35%-50% (increasing with compression rates) for females.

It can be therefore concluded that when using TEO-PWP features, the AMR compression reduced the AER accuracy by about 28%-33% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by up to 10% and made it flat across compression rates.

GP-T&GP-F (Figure 5.5)

Without hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech without HLS was about 55% for both genders.

For AMR without HLS, the AER accuracy from compressed speech on average lead to 40%-50% for males and flat 50% with increasing compression rate for females.

With hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech with HLS was about 51% for both genders.

For AMR with HLS, the average AER accuracy from compressed speech was about 20% for males and females.

It can be therefore concluded that when using GP-T&GP-F features, the AMR compression reduced the AER accuracy by up to 10% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 30% and made it flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 4%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

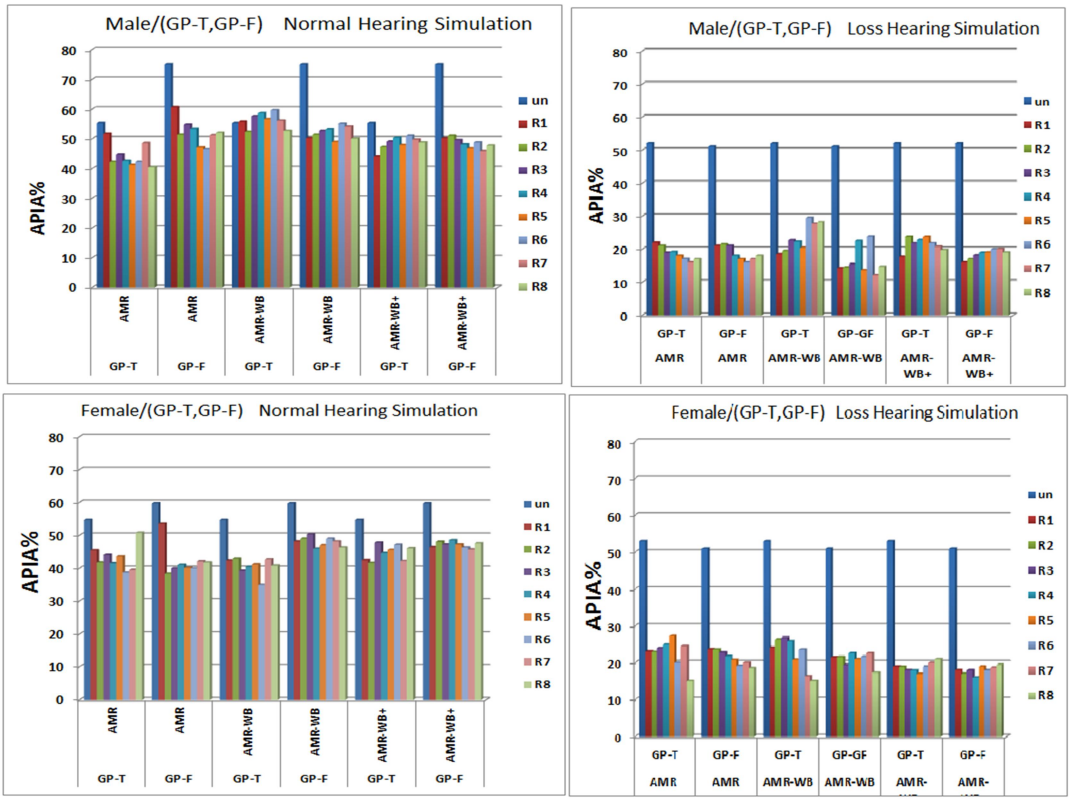

Figure 5.5: Average accuracy of multi-class emotion recognition for male and female speakers using GP-T&GP-F features; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order for both normal and hearing-loss simulation

5.4.2 Effect of AMR-WB and Hearing Loss Simulation on AER

MFCC (Figure 5.3)

Without hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech without HLS was about 72% for males and 75% for females.

For AMR-WB without HLS, the AER accuracy from compressed speech was constant at 60% for all rates and for both genders.

With hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech with HLS was about 52% for males and 55% for females.

For AMR-WB with HLS, the AER accuracy from compressed speech led to classification accuracy of 55% for males and 50% for females (almost flat across compression rates) for both genders.

It can be therefore concluded that when using MFCCs, the AMR compression reduced the AER accuracy by about 12%-15% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 10% - 15% and made it flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 20%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

TEO-PWP (Figure 5.4)

Without hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech without HLS was about 78% for both genders.

For AMR-WB without HLS, the AER accuracy from compressed speech lead to 78%-65% accuracy (decreasing with increasing compression rates) for both genders.

With hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech with HLS was about 52% for males and 55% for females.

For AMR-WB with HLS, the AER accuracy from compressed speech was about 50%-58% (increasing with compression rates) for males and flat 50% for females.

It can be therefore concluded that when using TEO-PWP features, the AMR compression reduced the AER accuracy up to 13% compared to uncompressed speech,

and the subsequently applied hearing loss simulation decreased this accuracy even further by about 20% and made it almost flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 23%-26%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

GP-T&GP-F (Figure 5.5)

Without hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech without HLS was about 75% for males and 60% for females.

For AMR-WB without HLS, the AER accuracy from compressed speech on average lead to 50%-60% for males and 40%-50% with increasing compression rate for females.

With hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech with HLS was about 50% for both genders.

For AMR-WB with HLS, the average AER accuracy from compressed speech was about 20%-30% for males and 20% for females.

It can be therefore concluded that when using GP-T&GP-F features, the AMR compression reduced the AER accuracy by up to 10%-25% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 20%-30% and made it almost flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 10%-25%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

5.4.3 Effect of AMR-WB+ and Hearing Loss Simulation on AER

In both MFCC (Figure 5.3)

Without hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech without HLS was about 72% for males and 75% for females.

For AMR-WB+ without HLS, the AER accuracy from compressed speech was constant at 55% for all rates for males and increasing from 50% to 55% with compression rates for females.

With hearing loss simulation (HLS)

When using the MFCC parameters, the AER accuracy from uncompressed speech with HLS was about 52% for males and 55% for females.

For AMR-WB+ with HLS, the AER accuracy from compressed speech led to classification accuracy of 42% (almost flat across compression rates) for males and 40%-42% for females.

It can be therefore concluded that when using MFCCs, the AMR compression reduced the AER accuracy by about 20% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 10% - 13% and made it flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 20%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

TEO-PWP (Figure 5.4)

Without hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech without HLS was about 78% for both genders.

For AMR-WB+ without HLS, the AER accuracy from compressed speech lead to 58% for males and 60% accuracy for females (almost flat across compression rates).

With hearing loss simulation (HLS)

When using the TEO-PWP parameters, the AER accuracy from uncompressed speech with HLS was about 52% for males and 55% for females.

For AMR-WB+ with HLS, the AER accuracy from compressed speech was about 40% (flat across compression rates) for both genders.

It can be therefore concluded that when using TEO-PWP features, the AMR compression reduced the AER accuracy up to 20% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 18%-20% and made it almost flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 25%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

GP-T&GP-F (Figure 5.5)

Without hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech without HLS was about 55%-75% for males and 55%-60% for females.

For AMR-WB+ without HLS, the AER accuracy from compressed speech on average lead to 50% flat across compression rates for both genders.

With hearing loss simulation (HLS)

When using the GP-T&GP-F parameters, the AER accuracy from uncompressed speech with HLS was about 50% for both genders.

For AMR-WB+ with HLS, the average AER accuracy from compressed speech was about 20% for males and 18%-20% for females.

It can be therefore concluded that when using GP-T&GP-F features, the AMR compression reduced the AER accuracy by up to 10%-20% compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 20%-32% and made it almost flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 5%-10%.

The degradation of AER accuracy due to compression was therefore significantly increased by simulation of hearing loss.

5.5 Conclusion

This experiment investigated effects of standard speech compression techniques on AER from speech modified in a way that simulates a typical hearing loss.

As indicated in Chapters 3 and 4, removal of high frequency components from speech spectrum leads to significant degradation of AER accuracy.

Given that, a typical age-related hearing loss is characterised by reduced ability to hear high frequency components of speech, hearing aids users may not be able to capture full emotional contents of speech and subsequently experience reduced ability to recognise emotions from speech signals.

This effect may be even larger when an impaired hearing person listens to compressed speech.

Experiments conducted in this chapter indicated that this hypothesis could be true. However, this is only an indication based on machine learning. The full proof would require subjective listening tests which are beyond the scope of this thesis.

It was shown that, for all types of feature parameters, compression methods and gender, the speech compression reduced the AER accuracy by about 10%-30% (depending on types of features and compression method) compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 5%-10% and made it flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 20%-30% depending on types of features and compression method.

Therefore, the degradation of AER accuracy due to compression was significantly increased by simulation of hearing loss. Generally no significant differences between genders in the above trends were observed.

Chapter 6: Effect of Speech Compression and Noise on AER

6.1 Preview

This chapter explores how the combined effects of white Gaussian noise and speech coding on the accuracy of AER from speech signals. As explained in Section 2.3, noise is an important factor that affects human understanding of both linguistic and paralinguistic aspects of speech signals [55], [56]. It is therefore likely that machine based recognition of emotions could be more challenging in the presence of noise. Given the fact that machine communications are also likely to use compressed speech, effects of both factors noise and compression on AER are investigated. Multi-class emotion and training were used in the classification process. The results showed that, the best performing under noisy conditions features were MFCCs and the best performing speech compression algorithms was AMR-WB.

6.2 Method

Speech signals are affected by noise in mobile phones, call centres, medical centres and police stations. As phone calls use speech compression, it is therefore important to understand how different noise levels affect AER from uncompressed and compressed speech.

6.2.1 Database

The emotion recognition experiments were conducted on the Berlin Emotional Speech (BES) database described in [9].

The database contains speech samples that represent seven emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]) with linguistically neutral content. The aim was to recognise emotions using only acoustic cues.

All files were available in wav audio format recorded at 16 kHz sampling rate and 16-bit amplitude resolution. Table 3.1 shows the number of available speech samples for different emotions.

6.2.2 Experimental Framework

As shown in Figure 6.1, standard signal classification pipeline has been applied.

In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.

In all experimental cases, the speech samples were first normalised into the range ±1.

After detection of voiced/silence, voiced speech frames were concatenated and used in the processing framework illustrated in Figure 6.1.

Three different levels of white Gaussian noise producing SNR=5dB, SNR=10dB and SNR=15dB were added to either compressed or uncompressed speech.

The speech was compressed using three different compression methods: AMR, AMR-WB and AMR-WB+ described in Chapter 3.

The female and the male speech samples were tested separately to determine the effect of gender on the classification results.

After addition of noise to either compressed or uncompressed speech, feature parameters were calculated.

The experiments used three types of feature parameters: MFCCs, TEO-PWP parameters, as well as, glottal time (GP-T) and frequency (GP-F) parameters.

The features were then applied to perform training leading to generation of acoustic class models representing anger, happiness, sadness, fear, disgust, boredom and neutral speech.

Acoustic class models were generated using the classical GMM modelling method with three mixtures [4].The GMM software was obtained from the HTK toolbox [54].

After generation of acoustic class models, the same process of noise addition and feature extraction was applied to test speech samples. The features were then passed to the Bayesian classifier typically associated with the GMM [68].

Classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10],

In each case of gender, noise level, type of features and compression method, a separate training and testing (classification) procedures were conducted [57].

In all experimental cases, the training and classification process was repeated 15 times, each time with different mutually exclusive training and testing sets selected using a stratified data-selection procedure [10].

For each repeat, 80% of data were used in training and 20% were used in testing.

Figure 6.1: Emotion speech recognition from noisy compressed or uncompressed speech. White Gaussian noise was added to either compressed or uncompressed speech to generate noisy signals with SNR values of 5dB, 10dB and 15dB.

6.2.3 Noise Addition

This section investigates the combined effects of noise and speech coding on AER [56].

To determine the effect of noise on AER from either compressed or uncompressed speech in a controlled way, different levels of white Gaussian noise were added to speech using Matlab function awgn() before feature extraction.

The framework for this experiment is illustrated in Figure 6.1.

In signal processing, white noise is a random signal with a constant power spectral density. In discrete time, white noise is a discrete signal whose samples are regarded as a sequence of uncorrelated random variables with zero mean and finite variance.

The awgn() allowed to specify particular values of the signal to noise ratio (SNR) as one of the input parameters.

The white Gaussian noise was added to either compressed or uncompressed speech in a way that the SNR was increasing in 5dB incremental steps, and generating noisy signals with SNR per sample equal to 5dB, 10dB and 15dB [55], [56].

The awgn() function [74] was setup to measure the actual power of the speech signals before adding the white noise.

Since the SNR parameter has been defined as a ratio of speech signal power to standard deviation of the white nose, the lower was the SNR value, the more noisy was the signal.

6.2.4 Calculation of Speech Features for Emotion Recognition

Acoustic speech features, including MFCC [32], TEO-PWP [19] and GP-T&GP-F domain parameters were used to differentiate between different emotions. All feature parameters were calculated on a frame-by-frame basis with a frame length of 256 samples and 50% overlap between frames.

MFCCs: MFCCs have been reported to provide good performance in speaker recognition and emotion classification in speech [15], [16], [17]. For each frame, the Fourier transform and the energy spectrum were estimated and mapped onto the mel-frequency scale. The DCT of the mel-log energies was estimated, and the first 12 DCT coefficients provided the MFCC values used in the modelling and classification process.

TEO-PWP: Features derived from the TEO [18], [53], [7] have been previously applied in emotion, stress and depression [20], [22] classification systems. Calculation of the TEO parameters involves estimation of the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through PWP decomposition [53], [24] into close estimates of the critical bands. For each

frame of length 256 samples, values of the TEO instantaneous energy of a given signal x[n] were calculated using Equation 3.2 [26].

Instantaneous energy was then used to evaluate the TEO autocorrelation function values using Equation 3.3 [18], where *M* was the number of samples in the given frame. After smoothing with cubic splines, the area under the autocorrelation contour was calculated for each frame within each of the 17 frequency bands.

GP-T&GP-F: Glottal features have been shown to provide efficient classification of emotion [33], [30] and depression [21], [22], [23] in speech. An IAIF algorithm based on DAP modelling was used to generate the glottal wave, and the glottal parameters were calculated using procedures included in the TKK Aparat Toolbox [27]. GP-T features were represented by nine different parameters describing amplitudes, timing and duration of the opening and closing phases of the vocal folds. GP-F features included three different parameters calculated from the spectrum of the glottal wave. These parameters described the differences between amplitudes of the first and second harmonic components of the glottal wave, the ratio of the sum of amplitudes of the higher harmonics to the amplitude of the first harmonic, and the spectral decay of the glottal waveform.

6.3 Discussion and Results

This section shows how the combined effects of speech compression (AMR, AMR-WB and AMR-WB+) and white Gaussian noise addition (SNR=15dB, SNR=10dB and SNR=5dB) to speech signals affects AER.

The experimental results for both genders and for three compression methods are presented in Figures 6.2–6.10.

The following sections describe and analyse the results separately for each SNR levels in an increasing level of noise (SNR=15dB, SNR=10dB and SNR=5dB).

6.3.1 Effect of 15dB Noisy Speech with Three Different Standard Speech Compression Techniques (AMR, AMR-WB and AMR-WB+)

The AER results for noisy speech with SNR=15dB are presented in Figures 6.2 (MFCCs), Figure 6.3 (TEO-PWP) and Figure 6.4 (GP-T&GP-F). Each figure shows

the percentage of average AER accuracy against compression bit rates. The bit rates increase (i.e. the amount of compression decreases) from left to right and the values in kbps are listed in Table 3.2.

Each picture shows three lines, where each line corresponds to different type of compression: AMR (blue), AMR-WB (red) and AMR-WB+ (green).

Classification based on MFCCs (Figure 6.2)

For uncompressed speech the AER accuracy under SNR=15% with MFCCs achieves about 72% for females and to 70% for males. This means that the accuracy is reduced only by about (1% for females and 3% for males) compared to clean speech without noise (Figure 3.2).

For MFCC parameters with SNR=15dB, the AMR-WB shows the highest performance of around (65% for both genders) accuracy almost constant across all compression rates. This is followed by AMR-WB+ (increasing from 55% to 60% with the bit rates for females and constant at 60% for males). Finally the worse performance is given by the narrow-band AMR (50% for and 52% for males females across all compression rates).

There are generally no significant gender dependent differences.

Classification based on TEO-PWP (Figure 6.3)

For uncompressed speech the AER accuracy under SNR=15% with TEO-PWP achieves about 50% for both genders. This means that the accuracy is reduced by about 29% for both genders compared to clean speech without noise (Figure 3.3).

For TEO-PWP parameters with SNR=15dB, the AMR-WB and AMR-WB+ show very similar performance with around 50% accuracy across all rates for both genders. The AMR compression drops down to 40% for female speakers (constant for all rates) but stays at 50% for male speakers (for all rates).

Gender differences have been observed only for AMR compression with male voices outperforming female voices.

Classification based on GP-T&GP-F (Figure 6.4)

For uncompressed speech the AER accuracy under SNR=15% with GP achieves about 72% for females and to 70% for males. This means that the accuracy is reduced only by about (1% for females and 3% for males) compared to clean speech without noise (Figure 3.2).

In general the GP parameters have shown very low performance (20% - 50% accuracy) under SNR=15dB for all compression rates.

The best performing compression method was AMR-WB with GP-T features leading to almost constant accuracy of 45% across all rates for female speakers and oscillating between 40% and 50% with compression rates for male speakers.

No significant gender differences have been observed.

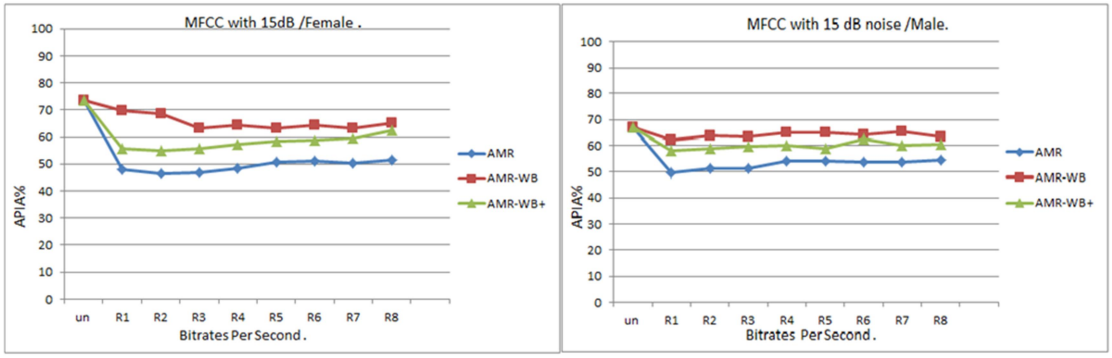

Figure 6.2: Average accuracy of multi-class emotion recognition for MFCC features with 15dB noise for females and male; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

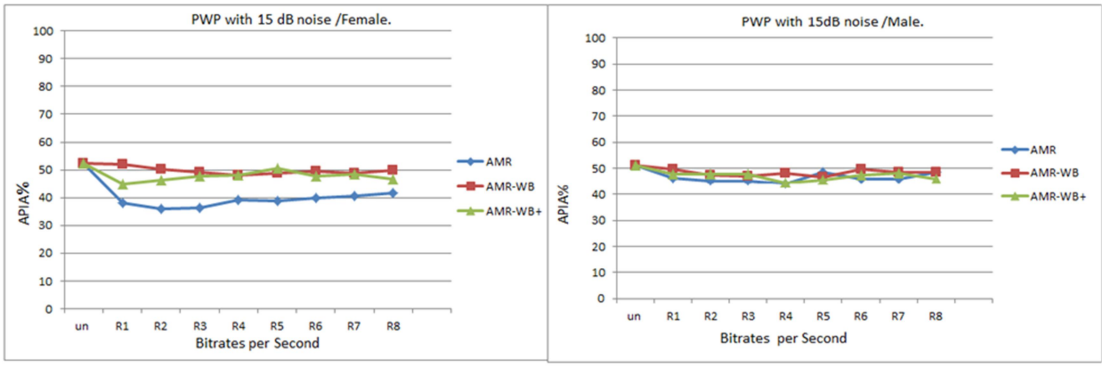

Figure 6.3: Average accuracy of multi-class emotion recognition for TEO-PWP features with 15 dB noise for females and males; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

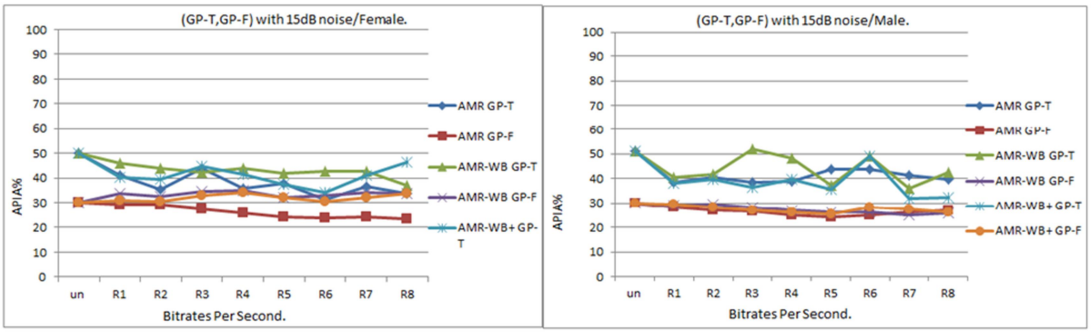

Figure 6.4: Average accuracy of multi-class emotion recognition for (GP-T, GP-F) features with 15 dB noise for females and males; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

6.3.2 Effect of 10dB Noisy Speech with Three Different Standard Speech Compression Techniques (AMR, AMR-WB and AMR-WB+)

The AER results for noisy speech with SNR=10dB are presented in Figures 6.5 (MFCCs), Figure 6.6 (TEO-PWP) and Figure 6.7 (GP-T&GP-F). Each figure shows the percentage of average AER accuracy against compression bit rates. The bit rates increase (i.e. the amount of compression decreases) from left to right and the values in kbps are listed in Table 3.2.

Each picture shows three lines, where each line corresponds to different type of compression: AMR (blue), AMR-WB (red) and AMR-WB+ (green).

MFCC (Figure 6.5)

For uncompressed speech the AER accuracy under SNR=10% with MFCCs achieves about 70% for females and to 65% for males. This means that the accuracy is reduced by about (2% for females and 10% for males) compared to clean speech without noise (Figure 3.2).

For MFCC parameters with SNR=10dB, the AMR-WB shows the highest performance of around (60% for females and 65% for males) accuracy almost constant across all compression rates. This is followed by AMR-WB+ (55% for females and 59% for males) and constant across bit rates. Finally the worse performance is given by the narrow-band AMR (50% for females, constant across rates, and 50%-55% for males females varying with rates).

Small gender differences were observed. Generally male voices outperformed female voices by about 5%..

TEO-PWP (Figure 6.6)

For uncompressed speech the AER accuracy under SNR=10% with TEO-PWP features achieves about 45% for females and to 50% for males. This means that the accuracy is reduced by about (34% for females and 29% for males) compared to clean speech without noise (Figure 3.3).

For TEO-PWP parameters with SNR=10dB, the AMR-WB and AMR-WB+ show very similar performance with around 50% accuracy across all rates for female speakers and about 50%-54% for male speakers. The AMR compression drops down to 40% - 45% for female speakers (constant for all rates) and stays at 48% for male speakers (for all rates).

Small gender differences were observed with male voices outperforming female voices by about 4%.

GP-T&GP-F (Figure 6.7)

For uncompressed speech the AER accuracy under SNR=10% with GP achieves about 35% for females and to 30% for males. This means that the accuracy is reduced by about (20%-40% for females and 10%-30% for males) depending on type of GP compared to clean speech without noise (Figure 3.4).

In general the GP parameters have shown very low performance (20% - 50% accuracy) under SNR=10dB for all compression rates.

For both genders, the best performing compression methods were AMR-WB and AMR-WB+ with GP-T features leading to accuracy oscillating between 40%-50% depending on compression rates.

No significant gender differences have been observed.

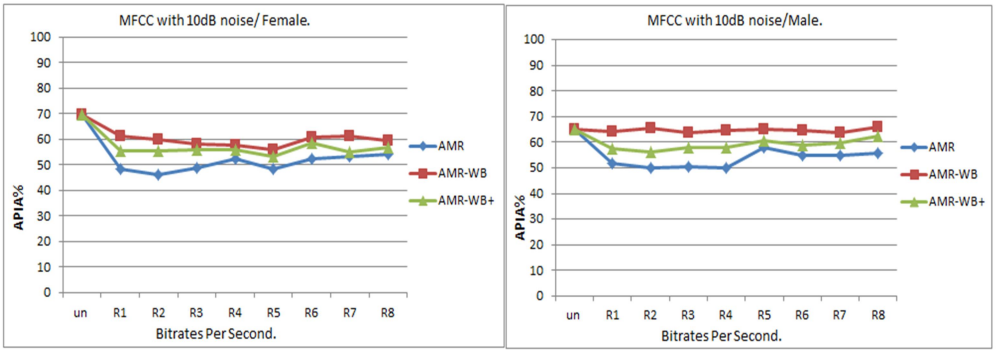

Figure 6.5: Average accuracy of multi-class emotion recognition for MFCC features with 10 dB noise for females and males; UN denotes uncompressed speech and R1–R8 are compression rates in increasing order

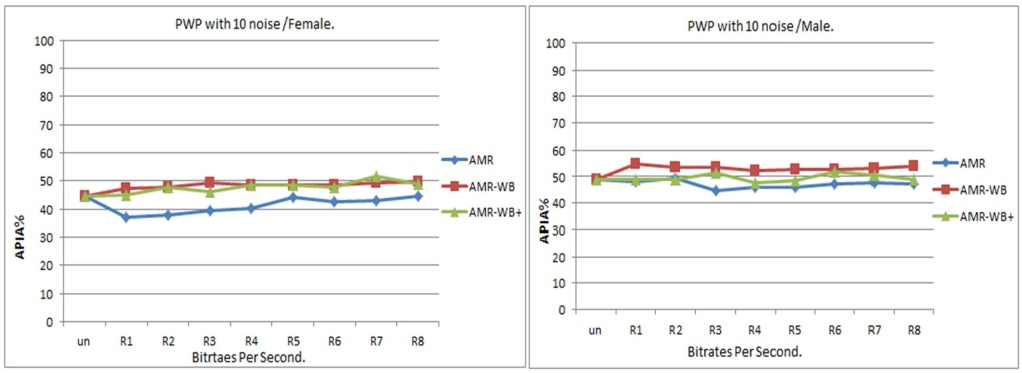

Figure 6.6: Average accuracy of multi-class emotion recognition for TEO-PWP features with 10 dB noise for females and males; UN denotes uncompressed speech and R1–R8 are compression rates in increasing order

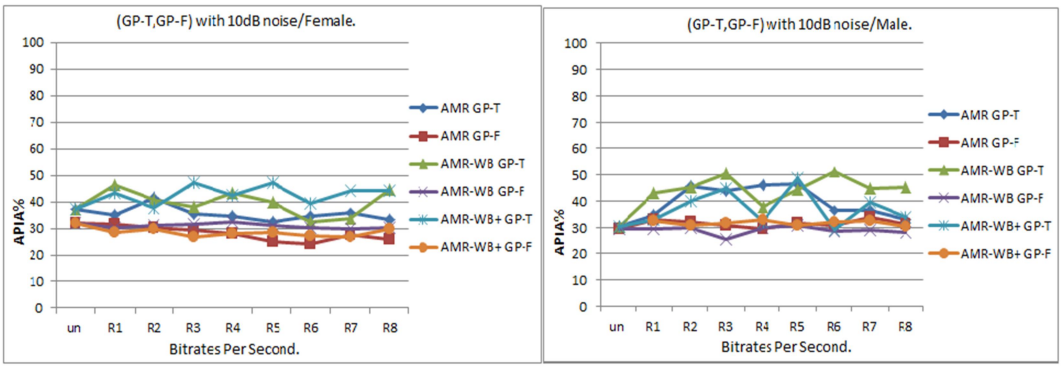

Figure 6.7: Average accuracy of multi-class emotion recognition for (GP-T, GP-F) features with 10 dB noise for females and males; UN denotes uncompressed speech and R1–R8 are compression rates in increasing order

6.3.3 Effect of 5dB Noisy Speech with Three Different Standard Speech Compression Techniques

The AER results for noisy speech with SNR=5dB are presented in Figures 6.8 (MFCCs), Figure 6.9 (TEO-PWP) and Figure 6.10 (GP-T&GP-F). Each figure shows the percentage of average AER accuracy against compression bit rates. The bit rates increase (i.e. the amount of compression decreases) from left to right and the values in kbps are listed in Table 3.2.

Each picture shows three lines, where each line corresponds to different type of compression: AMR (blue), AMR-WB (red) and AMR-WB+ (green).

MFCC (Figure 6.8)

For uncompressed speech the AER accuracy under SNR=5% with MFCCs achieves about 60% for females and to 65% for males. This means that the accuracy is reduced by about (12% for females and 10% for males) compared to clean speech without noise (Figure 3.2).

For MFCC parameters with SNR=5dB, the AMR-WB shows the highest performance of around (60% for both genders) accuracy almost constant across all compression rates. This is followed by AMR-WB+ (increasing from 50% to 60% with the bit rates for both genders). Finally the worse performance is given by the narrow-band AMR (50% for both genders across all compression rates).

No significant gender dependent differences were observed.

TEO-PWP (Figure 6.9)

For uncompressed speech the AER accuracy under SNR=5% with TEO-PWP features achieves about 45% for both genders. This means that for both genders, the accuracy is reduced by about 34% compared to clean speech without noise (Figure 3.3).

For TEO-PWP parameters with SNR=5dB, the AMR-WB and AMR-WB+ show very similar performance with around 45% accuracy across all rates for both genders. The AMR compression drops down to 35%-40% for female speakers (increasing slightly with bit rates) but stays at 45% for male speakers (for all rates).

Gender differences have been observed only for AMR compression, with male voices outperforming female voices by about 5%.

GP-T&GP-F (Figure 6.10)

For uncompressed speech the AER accuracy under SNR=5% with GP achieves about 30%-40% for females and to 30%-45% for males depending on type of GPs. This means that the accuracy is reduced by about (20%-30% for females and 45%-35% for males) depending on type of GP compared to clean speech without noise (Figure 3.4).

In general the GP parameters have shown very low performance (20%-40% accuracy) under SNR=5dB for all compression rates.

The best performing compression methods were AMR-WB and AMR-WB+ with GP-T features leading to accuracy of 40% across all rates for female speakers and oscillating between 40% and 50% with compression rates for male speakers.

Small gender differences were observed, with male voices outperforming female voices up to 10%.

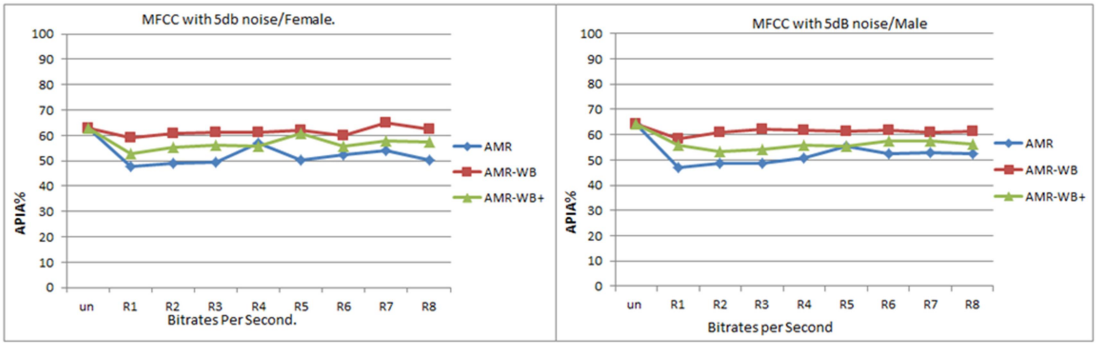

Figure 6.8: Average accuracy of multi-class emotion recognition for MFCC features with 5 dB noise for females and males; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

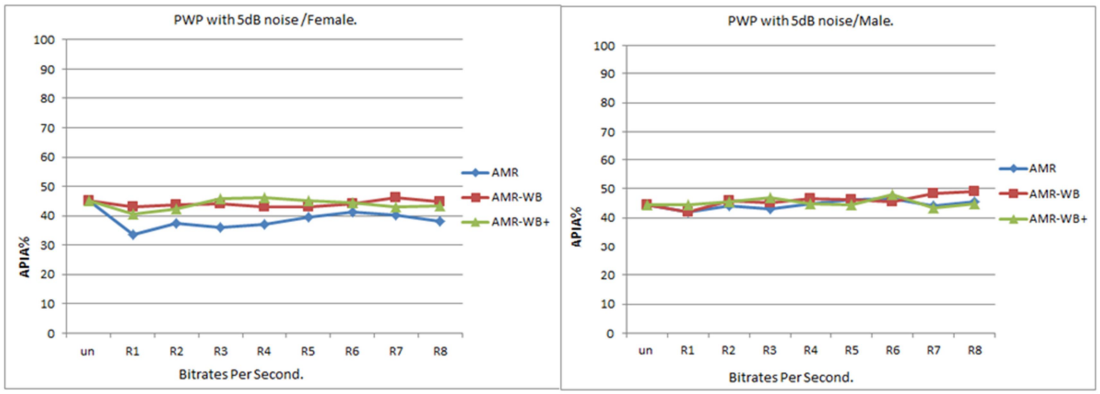

Figure 6.9: Average accuracy of multi-class emotion recognition for TEO-PWP features with 5 dB noise for females and males; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

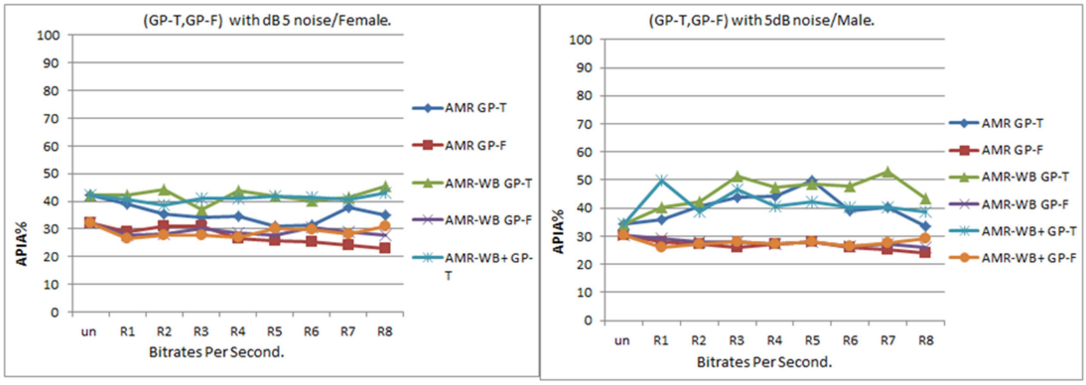

Figure 6.10: Average accuracy of multi-class emotion recognition for GP-T&GP-F features with 5 dB noise for females and males; Un denotes uncompressed speech and R1–R8 are compression rates in increasing order

6.4 Conclusion

Effects of speech compression and noise on the accuracy of AER from speech signals were investigated.

AER experiments aiming to simultaneously classify 7 different emotional states were conducted using three different types of features (MFCCs, TEO-PWP and GP-T&GP-F) and the GMM modelling/classification method. The experiments were conducted on publically available Berlin Emotional Speech research data.

White Gaussian noise with SNR=15dB, 10dB and 5dB was added to either uncompressed or compressed speech prior to classification. Three different speech compression methods were used; narrow band AMR, AMR-WB and AMR-WB+.

As it could be expected, addition of noise to either uncompressed or compressed speech reduced the accuracy of AER.

The amount of accuracy reduction depended on SNR, type of speech features, gender and compression method.

Gender differences were observed, with male voices generally outperforming female voices by 4%-10%. These differences become more prominent as the SNR values decrease. While for the AMR-WB and AMR-WB+ compression these differences are very small, they become clearly noticeable when the narrow-band AMR compression is applied.

These observations are consistent with previous reports of strong gender dependency in speech emotion [7], [19], [24], [33], stress [8], [18], [19], [20], [53] and depression classification [20], [21], [22], [23].

Higher resilience to noise observed in male voices can be attributed to the fact that male voices show generally lower values of fundamental frequency F0, as well as formant and harmonic frequencies compared to female voices. This means that (1) addition of high frequency noise may have lesser effect on male speech than female speech; (2) band reduction introduced by AMR compression may have smaller effect on male voices than female voices.

For uncompressed speech, the AER accuracy was reduced depending on the SNR value and type of speech feature parameters. For all values of SNR, the best performing features were MFCCs followed by the TEO-PWP, and the worse performing glottal parameters GP-T&GPF.

The same order of features' performance was observed for AER from compressed speech. Here again the MFCCs provided the highest performance followed by TEO-PWP and the worse performing GP-T&GP-F. The same order was holding for all values of SNR values and for both genders.

For all three SNR values (15dB, 10dB and 5dB) AER provided the highest accuracy values for the AMR-WB algorithm. The AMR-WB was followed by slightly worse performing AMR-WB+, and the least performing narrow-band AMR.

This means that the AMR-WB was found to be the best performing and therefore, the most robust or noise resilient speech compression method for AER.

High performance of the AMR-WB compression in AER can once again attributed to the fact that the algorithm preserves high bandwidth of speech which in terms implies that the emotional cues are most likely to be conveyed through high frequency components of speech signals.

If the high part of speech spectrum is preserved, addition of noise leads much smaller degradation of AER accuracy compared to cases when the high frequency information is either completely (AMR) or partially removed (AMR-WB+).

In conclusion the best performing under noisy conditions features were MFCCs and the best performing speech compression algorithms was AMR-WB.

Chapter 7: Effect of Speech Compression on AER Based on Speech Spectrograms

7.1 Preview

This chapter investigates the effects of standard AMR, AMR-WB, MP3 audio and AMR-WB+ speech codecs on AER based on speech spectrogram features. This investigation has been motivated by the fact that speech spectrograms have been reported to provide current state of the art performance in AER. The recognition process was based on two types of features including speech spectrograms (SS) and two types of spectrograms incorporating auditory perception criteria: speech spectrogram energy features derived from critical bands (SS-CB) and speech spectrogram energy features derived from bark bands (SS-Bark). The aim was to observe if application of state of the art features and subjective auditory criteria can reduce the detrimental effects of speech compression on AER. The results showed that, in general, all speech compression rates reduced emotion recognition accuracy. The SS-CB features provided more robust performance under various compression conditions and generally outperformed both the SS and the SS-Bark features. The SS-Bark parameters performed slightly better than the SS parameters. The amount of degradation due to compression varied across compression methods, compression rates and genders. The accuracy of emotion recognition using the AMR-WB codec was higher than that of AMR, AMR-WB+ and MP3. These results indicated that incorporation of auditory perceptive criteria into speech features can not improve AER performance. However it was observed that application of speech spectrogram features (SS) shows very robust performance under AMR-WB and AMR-WB+ compression reducing the AER accuracy only by a small amount (5%) compared to uncompressed speech.

7.2 Introduction

An investigation of effects of speech compression on AER based on speech spectrograms has been motivated by the fact that speech spectrograms have been

reported to provide current state of the art performance when applied as input vectors to Deep Neural Network (DNN) models [70]. Unlike high level features such as MFCCs, glottal or TEO parameters, the speech spectrograms require very little processing. The processing is limited to calculation of short-time Fourier transform and therefore can be performed in real-time. This computational simplicity of speech spectrograms reduces the AER system latency while maintaining complete set of paralinguistic (emotional) cues in a very compact form.

7.3 Method

7.3.1 Speech Data

The emotion recognition experiments were conducted on the Berlin Emotional Speech (BES) database described in [9]. The database contains speech samples that represent seven emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]) with linguistically neutral content. The sampling frequency of the speech samples was 8 kHz. Table 3.1 presents the number of available speech samples for the different emotions.

7.3.2 Speech Spectrograms

A two-dimensional magnitude spectrogram is a graphical display of the magnitude of the time-varying spectral characteristics of speech. It can be used to calculate numerous parameters such as energy, fundamental frequency (F0), formants and timing. These parameters are the acoustic features of speech most often used in automatic stress and emotion recognition systems [18]. Majorities of these systems analyze each parameter separately, and then combine them into a set of feature vectors. The approach presented here aimed to capture all of these characteristics at once, and preserve the important underlying dependencies between different parameters through analysis of speech spectrograms.

The speech spectrograms were previously not applied to the stress and emotion recognition problem, however other closely related applications have been reported. Kleinschmidt *et al.* [72], [58] applied a 2D Gabor filter bank to mel-spectrograms. The

resulting outputs of the Gabor filters were concatenated into one-dimensional vectors and used as features in the speech recognition experiments. Chih *et al.* [59] applied a similar method to the process speech discrimination and enhancement. In recent studies Ezzat *et al.* [60]-[63] described a spectro-temporal Gabor filter bank and used it to analyze localized patches of spectrograms, which showed advantages over one-dimensional features in word recognition. Meyer applied Gabor-shaped localized spectro-temporal features to successfully enhance automatic speech recognition performance [64].

As described in [45] when for example looking at the spectrograms of sentences pronounced by the same speaker under different emotional conditions it can be observed that different emotions are characterized by amplitude gradients and distributions of energy across frequencies. Moreover, the spectral energy decreases with frequency, however the rate of this decrease differs across different emotions. These observations indicate that speech spectrograms could be used to efficiently differentiate between different emotional states of speakers.

Using these observations a number of feature proposed in [64] were calculated to conduct AER experiments on compressed and uncompressed speech.

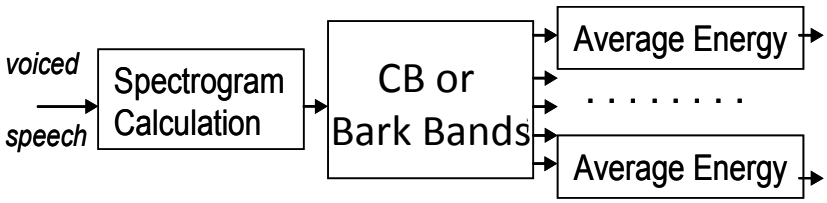

Figure 7.1: Features generation from the auditory frequency bands of spectrograms using different auditory scales (SS-CB, SS-Bark)

7.3.3 Calculation of Speech Spectrograms (SS)

Speech spectrograms (SS) were calculated using short-time Fourier analysis applied to 256-point frames of voiced speech with 50% overlap. The global maximum of the absolute magnitude was calculated for each spectrogram, and the absolute magnitude level at 50dB below the maximum value was chosen as the minimum and set to 0dB. All absolute magnitudes below the minimum level were also set to 0dB, and all absolute magnitudes between the minimum and maximum levels were mapped into

the range of 0dB-50dB. The 50dB value was determined experimentally as providing the best classification results.

7.3.4 Calculation of Speech Spectrogram Energy Features Derived from Critical Bands (SS-CB) and Bark Bands (SS-Bark)

The process of speech perception by the human auditory system shows high sensitivity to sounds occurring within specific frequency bands called the critical bands. The critical bands were introduced by Fletcher and Munsonin the 1940s, who referred to the frequency bandwidth of the then loosely defined auditory filter [25]. Critical bands were determined using simple listening tests. The listeners were presented with a pure tone submersed in white noise of a limited bandwidth. The amplitude of the tone was gradually decreased and the level was recorded when the listener could no longer hear it. The bandwidth of the noise was then reduced and the test was repeated. It was found, that the level where the listener was unable to hear the tone remained the same until the bandwidth of the noise was reduced to a critical width. Once the bandwidth of the noise was within this critical width, the listener's ability to hear the tone increased. It was concluded, that the auditory system works like a bank of filters. The widths of the filters are constrained to the critical points on either side of the tone and any noise outside this region is ignored. The critical points collected by Fletcher and Munson can be used as a scale to describe human hearing. The bands, defined by the region between the critical points, represent the bandwidths of the filters in the ear's psychological filter-bank. Fletcher and Munson named their bands, the critical bands.

Table 7.1 provides a list of lower and upper boundaries of the critical bands in Hz within the speech bandwidth ranging from 0 to 4 KHz. The widths of the critical bands increase logarithmically with frequency and the centre frequencies Bc in Hz are equally distant on the log scale.

Zhou's [18], demonstrated that the extraction of characteristic features based on critical bands was an important factor increasing the correct classification rates in an automatic stress classification.

Since the work of Fletcher and Munson other types of auditory scales have been developed. Two of the most popular scales are the bark scale and the equivalent rectangular bandwidth scale.

The Bark scale is a psychoacoustic scale proposed by Eberhard Zwicker in 1961 [25]. The Bark scale represents critical bands rates given by a parameter z in Barks.

For a given frequency f in Hz, the corresponding values of z in Barks can be calculated as follows:

$$z = \left[26.81/(1+1960/f)\right] - 0.53 \quad \textbf{(7.1)}$$

The inverse operation is given as:

$$f = 1960/\left[26.81/(z+0.53)-1\right] \quad \textbf{(7.2)}$$

The Bark scale bands B_{Bark} in Hz can be then calculated for different values of z as follows:

$$B_{Bark} = \frac{52548}{z^2 - 52.56z + 690.39} \quad \textbf{(7.3)}$$

The lower and upper edges of the Bark scale bands within the range 0 to 4kHz are listed in Table 7.1.

Since the hearing system performs a temporal analysis that contributes to frequency resolution for low frequencies, auditory frequency resolution cannot be fully represented on the basis of z alone.

It has been postulated that, the auditory frequency resolution is better described by the equivalent rectangular bandwidth (ERB) [65], [66]. The equivalent rectangular bandwidth is a measure of auditory frequency bands, which approximates the auditory system as a bank of rectangular band-pass filters. The ERB bandwidth values in Hz for a given center frequency f in Hz can be calculated as:

$$B_{ERB} = 6.23 \cdot 10^{-6} f^2 + 9.339 \cdot 10^{-2} f + 28.52 \quad \textbf{(7.4)}$$

The lower and upper edges of the ERB scale bands within the range 0 to 4kHz are listed in Table 7.1.

The 2D spectrograms of energy spectral density (squared magnitudes) were divided into sub-bands based on three different auditory scales: critical bands, Bark scale, and ERB scale. For each sub-band a single value of the average energy $\hat{E}_i$ (i=1,...,N) was calculated using:

$$\hat{E}_i = \frac{1}{N_f N_t} \sum_{y=1}^{N_f} \sum_{x=1}^{N_t} s(x, y) \quad \textbf{(7.5)}$$

Where s(x,y) are the spectrogram values (squared magnitudes) at the time coordinates x and frequency coordinates y, N_f is the total number of frequency coordinates, N_t is the total number of time coordinates, and N is the total number of frequency bands (N=16 for critical bands, N=17 for the Bark scale and N=27 for the ERB scale [73]).

Table 7.1: Critical and Bark Bands

No	Critical Bands[Hz]		Bark Bands [Hz]	
	Lower	Upper	Lower	Upper
1	100	200	0	80
2	200	300	80	160
3	300	400	155	245
4	400	510	250	350
5	510	630	345	455
6	630	770	450	570
7	770	920	565	695
8	920	1080	700	840
9	1080	1270	830	990
10	1270	1480	990	1170
11	1480	1720	1170	1370
12	1720	2000	1365	1595
13	2000	2320	1590	1850
14	2320	2700	1850	2150
15	2700	3150	2145	2495
16	3150	3700	2505	2915
17			2910	3410

The resulting feature values were then concatenated into 1D vectors, and passed to the GMM classifier for modeling and classification. The flow chart of the feature extraction process is illustrated in Figure 7.1.

The speech samples representing either compressed or uncompressed speech were normalised into the range ±1. After the removal of noise and detection of voiced/ silence, voiced speech frames were concatenated and used in the two-stage processing shown in Figure 7.2. In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.

For both compressed and uncompressed speech, and for each feature/classifier combination, the training and classification process was run 15 times, each time with different training and testing sets selected using a stratified training and testing data-selection procedure [10]. For each run, 80% of the data were used in the training process and 20% were used in the testing. The classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10], where N_C is the number of test inputs correctly identified, N_T is the total number of test inputs and N_r is the number of repeated tests. The emotion recognition was tested for each gender separately. Table 3.2 shows the compression bit rates tested in the experiments. Note that the compression rates corresponding to R1–R8 differ between the different types of codecs. This needs to be taken into account when evaluating the experimental results described in Section 7.4.

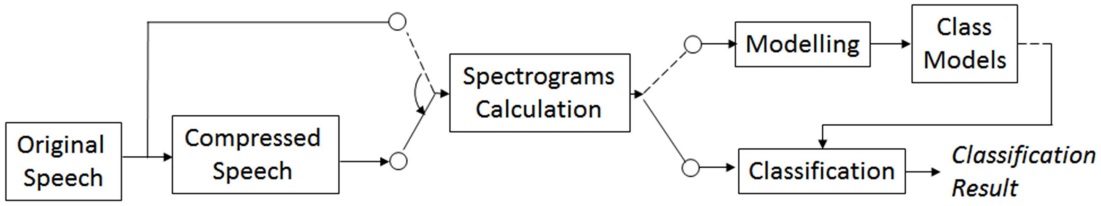

Figure 7.2: Framework of AER experiments using spectrogram features

7.3.5 Speech Compression Methods

The same speech compression techniques (AMR, AMR-WB, AMR-WB+ and MP3) as those used in Chapter 3 were applied. Descriptions of these techniques as well as, the corresponding transmission bit rates can be found in Chapter 3 and Table 3.2 respectively.

7.4 Results and Discussion

7.4.1 Classification Outcomes for Uncompressed Speech

The classification results for uncompressed speech are presented in Figure 7.3 (for SS parameters), in Figure 7.4 (for SS-CB parameters) and in Figure 7.5 (for SS-Bark parameters).

No significant differences between genders were observed.

For both genders, the AER based on speech spectrogram features (SS) under AMR-WB conditions lead to around 90% classification accuracy for males and 80% for females.

The SS outperformed both the SS-CB parameters (60%) and the SS-bark parameters (66%).

Thus, for uncompressed speech the SS-image parameters provided the best accuracy of recognition. Application of features related to human auditory perception lead to significant reduction of AER accuracy.

This outcome indicates that on its own, the spectral energy of critical or bark bands may not contain sufficient information needed to distinguish between different emotions.

A comparison with feature parameters used with uncompressed speech in Chapter 3 (Section 3.10.1) shows that the SS features with 96% accuracy clearly outperformed the MFCC parameters (73%), the TEO-PWP (78%), the GP-F (75%) and the GP-T parameters (55%). This is consistent with reports showing the state of teh art performance of the SS parameters [70].

7.4.2 Effect of the Narrow-Band AMR Compression on Emotion Classification

The classification results for AER based on AMR-compressed speech are presented in Figure 7.3 (for SS parameters), in Figure 7.4 (for SS-CB parameters) and in Figure 7.5 (for SS-Bark parameters).

Significant differences across genders were observed with male voices showing better performance than female voices for SS and SS-CB feature parameters however; in the case of SS-Bark features the female voices were slightly better.

For SS-image parameters, AMR compression led to slightly lower classification accuracy compared to uncompressed speech.

There was a decrease in classification accuracy 89% to 80% for males and from 85% to 79% for females, with compression bit rates decreasing from R8 (12.2 kbps) to R1 (4.75 kbps).

For SS-CB features, AMR led to classification accuracy decreasing from 50% to 40% for males and from 55% to 45% for females with the increase of compression bit rates from R8 (12.2 kbps) to R1 (4.75 kbps).

For SS-Bark features, AMR led to classification accuracy decreasing from 45% to 40% for males and from 58% to 45% for females with the increase of compression bit rates from R8 (12.2 kbps) to R1 (4.75 kbps).

These results show that like for an uncompressed speech, the SS features show the best performance for AMR-compressed speech outperforming both SS-CB and SS-bark parameters for both genders.

A comparison with feature parameters used with AMR-compressed speech in Chapter 3 shows that the SS features with their 79%–70% of average accuracy outperformed the MFCC parameters (40%–51%), the TEO-PWP (50%), the GP-F (60% for males and 52% for females) and the GP-T parameters ((51% for males and 45% for females)) working under the same AER conditions.

7.4.3 Effect of AMR-WB Compression on Emotion Classification

The classification results for AER based on AMR-WB-compressed speech are presented in Figure 7.4 (for SS parameters), in Figure 7.4 (for SS-CB parameters) and in Figure 7.5 (for SS-Bark parameters).

Significant differences across genders were observed with male voices showing better performance than female voices for SS, SS-CB andn SS-Bark feature parameters.

For SS-image parameters, AMR-WB-compression led to very small decrease of classification accuracy compared to uncompressed speech. The degradation affected mostly female speech, while male speech provided performance comparable with uncompressed speech.

Almost constant classification accuracy of about 90% across all compression rates was observed for male speakers and about 80% for female speakers.

For SS-CB features, AMR-WB compression led to classification accuracy decreasing from 65% to 60% for male speakers with the increase of compression bit rates from R8 (12.2 kbps) to R1 (4.75 kbps). However for female speakers the accuracy showed a constant value of about 60% across all compression rates.

For SS-Bark features, AMR led to classification accuracy decreasing from 65% to 55% for males with the increase of compression bit rates from R8 (12.2 kbps) to R1 (4.75 kbps). However for females, the accuracy showed a constant value of about 60% across all compression rates.

These results show once again that, like for an uncompressed speech, the SS features show the best performance for AMR-compressed speech outperforming both SS-CB and SS-bark parameters for both genders.

A comparison with feature parameters used with AMR-WB-compressed speech in Chapter 3 shows that the SS features with their 90% of average accuracy outperformed the MFCC parameters (60%), the TEO-PWP (67%-74%), the GP-F (55% -52%) and the GP-T parameters (40% - 48%) working under the same AER conditions.

It can be observed that the wide-band conditions given by the AMR-WB codec make the AER results almost independent on the compression bit rates.

Although, the AMR-WB AER results are still below the uncompressed speech levels, they are clearly higher compare to the narrow band AMR. This is clearly indicating that preservation of full speech bandwidth is very important for maintaining high accuracy of AER.

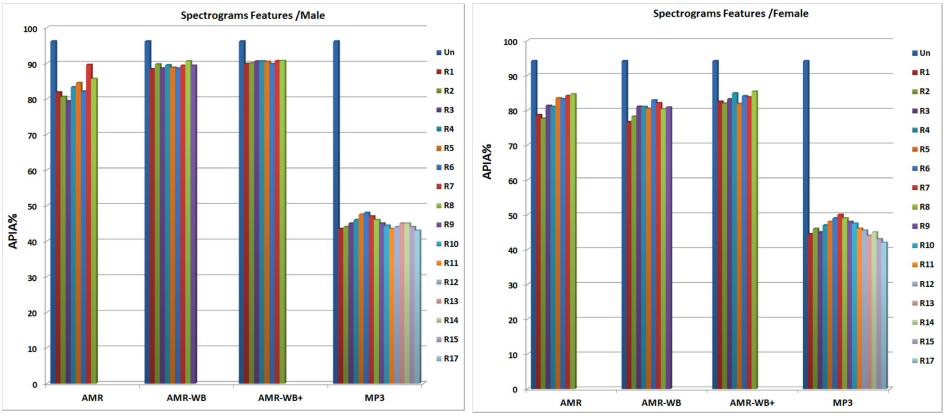

Figure 7.3: Average accuracy of multi-class emotion recognition for males and females using SS-image features; Un - denotes uncompressed speech and R1–R17 are bit rates in an increasing order

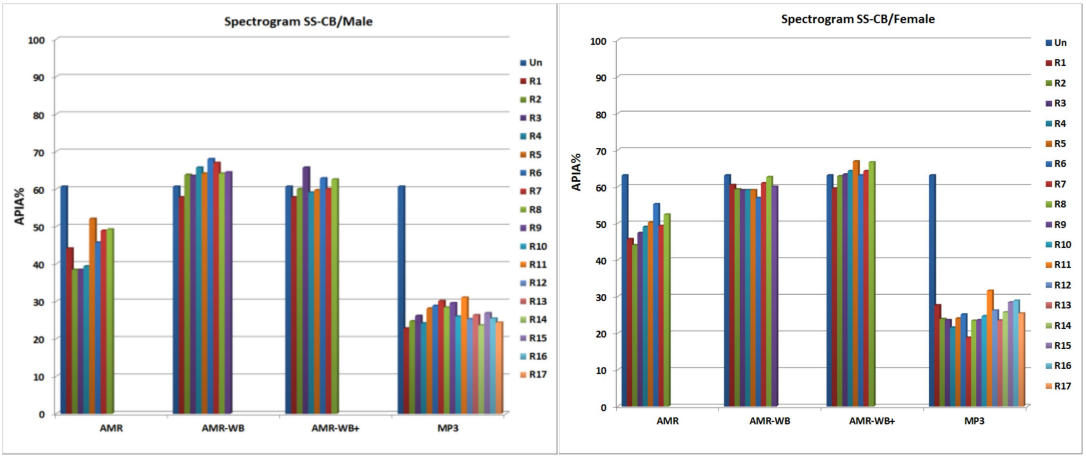

Figure 7.4: Average accuracy of multi-class emotion recognition for males and females using SS-CB features; Un denotes uncompressed speech and R1–R17 are bit rates in increasing order

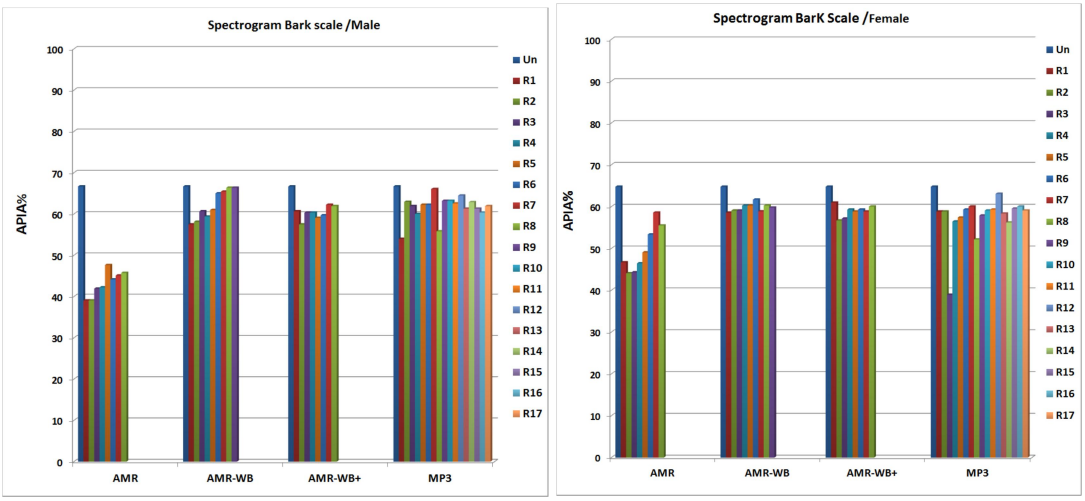

Figure 7.5: Average accuracy of multi-class emotion recognition for males and females using SS-Bark features; Un denotes uncompressed speech and R1–R17 are bit rates in increasing order

7.4.4 Effect of AMR-WB+ Compression on Emotion Classification

The classification results for AER based on AMR-WB+-compressed speech are presented in Figure 7.4 (for SS parameters), in Figure 7.4 (for SS-CB parameters) and in Figure 7.5 (for SS-Bark parameters).

Generally no significant differences across genders were observed.

For SS-image parameters, AMR-WB+-compression led to very small decrease of classification accuracy compared to uncompressed speech.

Almost constant classification accuracy of about 90% across all compression rates was observed for male speakers and about 84% for female speakers.

For SS-CB features, AMR-WB+ compression led to constant classification accuracy of 60% for male speakers across all bit rates. Similarly, for female speakers there was a constant accuracy of about 63% across all bit rates.

For SS-Bark features, AMR-WB+ led to constant classification accuracy of 60% for males and 63% for females across all bit rates.

Once again, like for an uncompressed speech, the SS features showed the best performance for AMR-WB+-compressed speech outperforming both SS-CB and SS-bark parameters for both genders.

A comparison with feature parameters used with AMR-WB-compressed speech in Chapter 3 shows that the SS features with their 90% of average accuracy outperformed the MFCC parameters (50%-55%), the TEO-PWP (58%-60%), the GP-F (48%) and the GP-T parameters (45% - 50%) working under the same AER-WB+ conditions.

Similar to AMR-WB, the results for AMR-WB+ were almost independent on the compression bit rates however not as high as those for AMR-WB.

7.4.5 Effect of MP3 Speech Compression on Emotion Classification

The classification results for AER based on MP3-compressed speech are presented in Figure 7.4 (for SS parameters), in Figure 7.4 (for SS-CB parameters) and in Figure 7.5 (for SS-Bark parameters).

Generally no significant differences across genders were observed.

The MP3 lead to very significant drop (30%-40%) in the AER accuracy for all features (SS, SS-CB and SS-Bark) compared to uncompressed speech as well as all other compression methods (AMR, AMR-WB and AMR-WB+).

For SS-image parameters, MP3 compression led to almost constant 45% accuracy across all bit rates. Indicating very significant decrease of classification accuracy compared to uncompressed speech. This reduction was similar for both genders.

For SS-CB features, MP3 compression led to almost constant classification accuracy of 27% for male and female speakers across all bit rates.

Similarly, for SS-Bark features, MP3 compression led to constant classification accuracy of about 62% for male and female speakers across all bit rates.

The SS features showed the best performance outperforming both SS-CB and SS-bark parameters for both genders.

Similar to AMR-WB and AMR-WB+, the results for MP3 were almost independent on the compression bit rates however significantly lower.

7.5 Conclusion

The most important conclusions coming from this experiment is that, speech features representing full spectrogram (SS) show only very small (5%) degradation of AER accuracy when applied to speech compressed with AMR-WB and AMR-WB+.

The very high performance of the SS features was consistently observed across all tested compression rates.

These results are generally consistent with previous reports describing state of the art performance of speech spectrograms. However this study for the first time shows very robust performance of the spectrogram features under speech compression conditions.

Incorporation of human auditory perception characteristics in the form of time-frequency features representing an average energy of either Critical Bands (SS-CB) or Bark Bands (SS-Bark) was found ineffective. It did not lead to an improvement of AER from compressed speech.

Both SS-CB and SS-Bark features were outperformed by the SS features representing full speech spectrograms indicating that spectral energy features alone are not sufficient for AER and have to be supported by richer information given by complete spectrograms.

The AMR-WB compression technique lead to the smallest degradation of the AER accuracy indicating that other techniques degraded the performance either due to severe bandwidth reduction (AMR) or optimization for music rather than speech (AMR-WB+ and MP3).

MP3 was found to be particularly bad in preserving emotional speech aspect leading to very big (30%-40%) reduction of AER accuracy compared to uncompressed speech.

Observed gender differences in most cases lead to higher AER accuracy for male voices than for female voices. This effect can possibly attributed to the fact that for male voices the fundamental frequency and formant frequencies are positioned at lower frequency ranges than for female voices. This means that speech compression methods which reduce high frequency contents of speech are more likely to remove vital high frequency harmonics of F0 as well as high frequency formants. As indicated

in Chapter 4, these high frequency components are likely to play important role in conveying emotional aspect of speech.

Chapter 8: Artificial Bandwidth Extension to Improve AER from Narrow-Band Coded Speech

8.1 Preview

Narrow-band speech coding techniques were previously found to reduce the accuracy of automatic AER, as well as speech and speaker recognition rates. ABE based on spectral folding and spectral envelope estimation has been applied to compressed narrow-band speech to test whether AER can be improved.

The modelling and classification of speech was performed with a benchmark approach based on the GMM classifier and a set of speech acoustic parameters including MFCCs, TEO and glottal parameters.

The tests used the BES database. In general, ABE led to an improvement in AER accuracy; however, the amount of improvement varied between different features, genders and speech compression rates.

In all cases, AER accuracy with ABE was at least 10% lower than for uncompressed speech.

8.2 Introduction

While modern mobile phone technologies can support a wide speech bandwidth, the existing mobile network infrastructure operates predominantly within narrow-band limitations. To maintain compatibility with analogue telephones, digital transmission adopted an 8 kHz sampling frequency, and speech signals were limited to a narrow-band range of 200 Hz to 3.4 kHz [122].

While this narrow bandwidth was adequate for speech intelligibility, it resulted in unnatural low-quality speech and was not necessarily adequate for the preservation of paralinguistic aspects of speech, such as emotions.

The AMR-WB and AMR-WB+ coding techniques [11], [13] were developed recently as the first significant improvement in the quality and intelligibility of telephone

speech. The sampling frequency was increased to 16 kHz, resulting in a wideband range of 50 Hz to 7 kHz. However, a complete wideband telephone network requires that all factors involved in the transmission chain support wideband characteristics; this includes the transmission network and the terminal devices, resulting in investments from network operators and end users.

The current speech transmission system is a mixture of traditional narrow-band and wideband terminals. The long transition period from a narrow-band to wideband system motivates the development of ABE methods, where the missing spectral content can be estimated from the narrow-band speech signal without modification to the current systems.

As indicated in [38], narrow-band speech coding techniques used by existing mobile phone networks lead to significant reductions in the accuracy of machine-based AER. This is largely because some of the speech codecs reduce the speech bandwidth and thus remove vital high-frequency information needed for differentiating between different emotions.

For the first time, this study investigated the concept of using an ABE technique to improve the accuracy of machine-based AER from narrow-band speech.

The purpose was not to challenge the current state-of-the-art technology in AER, but to determine whether the detrimental effects of bandwidth reduction on AER can, to some extent, be compensated for with ABE. Therefore, a standard AER approach and a standard emotion recognition database were used. If successful, the ABE methodology could facilitate the improvement of both linguistic and paralinguistic levels of speech recognition during the transition period from hybrid to a full wideband network environment.

8.3 Method

8.3.1 Narrow-Band Speech Compression

As indicated in [38], standard speech coding (or compression) techniques such as AMR, AMR-WB and AMR-WB+ reduced the overall emotion recognition accuracy when compared with uncompressed speech; however, the narrow-band AMR led to

the highest degradation of 15%–25% depending on gender and type of acoustic features. The AMR method was therefore used here to reduce the speech bandwidth before the ABE compensation method was applied. The AMR audio codec is an audio compression format optimised for speech coding. It consists of a multi-rate narrow-band speech codec that encodes narrow-band (200–3400 Hz) signals at variable bit rates ranging from 4.75 to 12.2 kbps, with toll quality speech starting at 7.4 kbps [67]. AMR was adopted as the standard speech codec by 3GPP in October 1999 and is now widely used in GSM, standard and UMTS—a third-generation mobile cellular system for networks.

The AMR is a hybrid speech coder [11], [12]; as such, it transmits both speech parameters and a waveform signal. LPC is used to synthesise the speech from a residual waveform. The LPC parameters are encoded as LSPs. The residual waveform is coded using ACELP. The AMR uses link adaptation to select from one of eight different bit rates based on link conditions (R1=4.75, R2=5.15, R3=5.9, R4=6.7, R5=7.4, R6=7.95, R7=10.2 and R8=12.2 kbps).

The speech was coded frame by frame with a frame size of 20 ms (160 speech samples at 8 kHz sampling rate). For each speech frame, the speech signal was analysed using an LP of order 10 to calculate the LP coefficients, adaptive codebook, fixed codebook parameters and the gains. Each frame was divided into sub-frames, and the mode was switched between subsequent sub-frames [122].

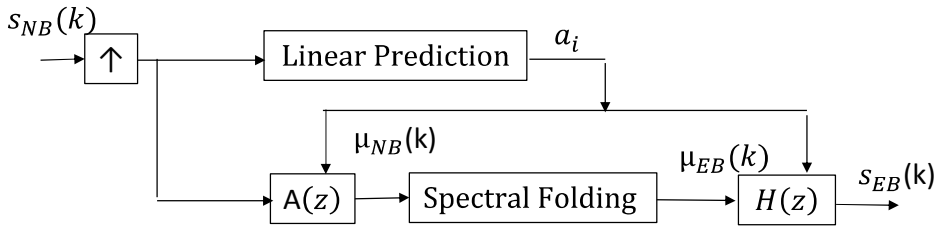

Figure 8.1: Bandwidth extension procedure. s_{NB} denotes narrow-band speech signal, μ_{NB} - narrow-band excitation signal, s_{EB} -extended-band speech signal, μ_{EB} - extended-band excitation signal, and a_i. are the vocal tract filter coefficients with *i*=1,...,15.

8.3.2 Artificial Bandwidth Extension

Figure 8.1 illustrates the general framework of the bandwidth extension process. The following sections describe the steps included in this framework.

8.3.3 Signal Up-Sampling

An important step required as an initial step in the process of bandwidth extension was an increase in the sampling rate of the input narrow-band speech signal to avoid aliasing after the band extension. Therefore, prior to bandwidth extension, the sampling rate of the speech signal was increased from 8 kHz to 16 kHz. The up-sampling was performed using cubic spline interpolation of the speech time waveform followed by low-pass filtering to remove high-frequency artefacts.

8.3.4 Bandwidth Extension Using Source-Filter Model

The artificial bandwidth extension was applied to the AMR-NB coded speech signal $s_{NB}(k)$ using an approach based on the linear source-filter model of the speech production [68], [44], [45] process, where the human vocal tract was modelled by an autoregressive filter *H(z)* given as:

$$\boldsymbol{H(z) = \frac{1}{A(z)} = \frac{1}{\sum_{i=0}^{N} a_i z^{-i}}} \quad \textbf{(8.1)}$$

The analysis (or inverse) filter *A(z)* in Equation 8.1 was given as:

$$\boldsymbol{A(z) = \sum_{i=0}^{N} a_i z^{-i}} \quad \textbf{(8.2)}$$

The filter coefficients $a_{NB}(k)$ were estimated using the LP method of order 15 [68], [44], [46]. Given an appropriate excitation signal $\mu(k)$ as an input, the filter would generate an estimate of the speech signal as an output. It could therefore be used to extrapolate the spectral envelope of the speech signal beyond the limits of the existing bandwidth.

Based on this model, the two-stage ABE technique illustrated in Figure 8.1 was implemented. In the first stage, a spectral extension of the excitation signal at the lower spectral end was performed using the spectral folding method. In the second

stage, the spectral envelope at the high-frequency end of the spectrum was extended using the source-filter model.

The second stage started with LP analysis applied to the narrow-band speech to estimate the narrow-band vocal tract filter coefficients $a_{NB}(k)$. These coefficients were used to construct the analysis filter $A(z)$, which was then applied to the narrow-band speech signal to generate the narrow-band excitation signal $\mu_{NB}(k)$ [44]–[46], [68].

After extending the spectrum of the excitation signal at the high-frequency range using spectral folding, a new extended-band (EB) excitation signal $\mu_{EB}(k)$ was generated. This signal was used as an input to the synthesis filter $H(z)$ to generate the EB speech signal $s_{EB}(k)$.

Due to the speech signal being considered stationary only within short periods of time, the ABE algorithms were implemented on a frame-by-frame basis. Each frame was generated using a 20 ms hamming window with 50% overlap between frames.

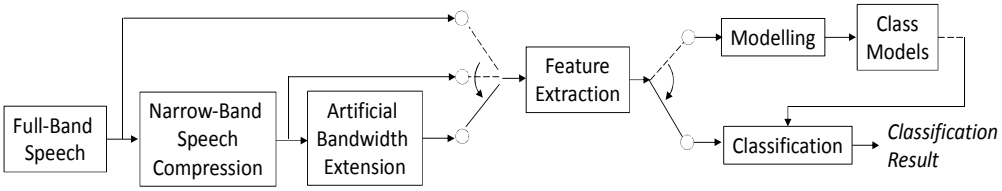

Figure 8.2: Experimental framework; comparing emotion recognition from speech with full band, narrow band and artificially extended band

8.3.5 Bandwidth Extension Using Spectral Folding

The spectrum of the excitation signal $\mu_{NB}(k)$ was extended by adding to this signal an AM signal with the carrier frequency Ω_M, as given in Equation 8.3:

$$\boldsymbol{\mu_{EB}(k) = \mu_{NB}(k) + \mu_{NB}(k)cos(\Omega_M k)} \quad \textbf{(8.3)}$$

By adding the AM signal, a mirror copy of the original signal narrow-band spectrum centred at the carrier frequency Ω_M = 4 kHz was added [3], [28].

8.3.6 Removal of Processing Artefacts

Given that the LP analysis and synthesis filters were not exactly mutually inverse, it was necessary to introduce a gain correction procedure to avoid spectral discontinuities between the original NB speech signal and the EB speech estimate. The correction gain was estimated as the ratio of the original NB signal gain and the gain of the estimated signal.

In addition to the gain changes, delays between the original and estimated signal were produced due to the action of the different filters. An additional delay filter was used to compensate for these delays.

8.4 Experiments and Results

8.4.1 Conversational Data

AER experiments were conducted on the BES database. Details of the BES data collection and validation procedures have been described [9]. The database contained speech samples representing seven categorical emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (five females and five males) fluent in German. Each speaker simulated all seven emotions while pronouncing 10 different utterances (five short [2–4 seconds] and five long [5–9 seconds]). The text of each utterance was designed to be emotionally neutral and thus provide no linguistics cues about its emotional content. The aim was to recognise emotions using acoustic cues only. All files were available in wav audio format recorded at 16 kHz sampling rate and 16-bit amplitude resolution. Table 3.1 provides the numbers of the available speech samples for different emotions.

8.4.2 Speech Emotion Recognition

A standard AER methodology [69] was used with the original full-band speech, narrow-band speech and speech with artificially extended bandwidth. The speech samples were normalised into the range ±1. After the removal of noise and detection of voiced/silence, voiced speech frames were concatenated and used in the processing framework illustrated in Figure 8.2. In the first stage (modelling), characteristic features representing known emotions were used to train the emotional class models.

In the second stage (classification), characteristic features from speech samples of unknown classes were compared with the models to determine the closest-matching emotional class.

For both compressed and uncompressed speech, and for each feature/classifier combination, the training and classification process was run 15 times, each time with different training and testing sets selected using a stratified training and testing data-selection procedure [10]. For each run, 80% of the data were used in the training process and 20% were used in the testing. The classification results were assessed using the average percentage of identification accuracy (APIA %) given in Equation 3.1 [10], where N_C was the number of test inputs correctly identified, N_T was the total number of test inputs and N_r was the number of repeated tests.

The AER task was set up to simultaneously recognise seven different emotions using BES data. All tests were conducted separately for each gender.

Three different types of features (MFCC, TEO-PWP and GP-T&GP-F) [38], [2], [82], [19] were tested. The feature vectors were calculated on a frame-by-frame basis with frame length of 256 samples and 50% overlap between frames.

The modelling and classification tasks were achieved using a third-order GMM algorithm integrated with the Bayesian classification decision procedure [28]–[31], [9], [33], which determined the most probable classes for given query samples. These algorithms were obtained from the HTK toolbox [54].

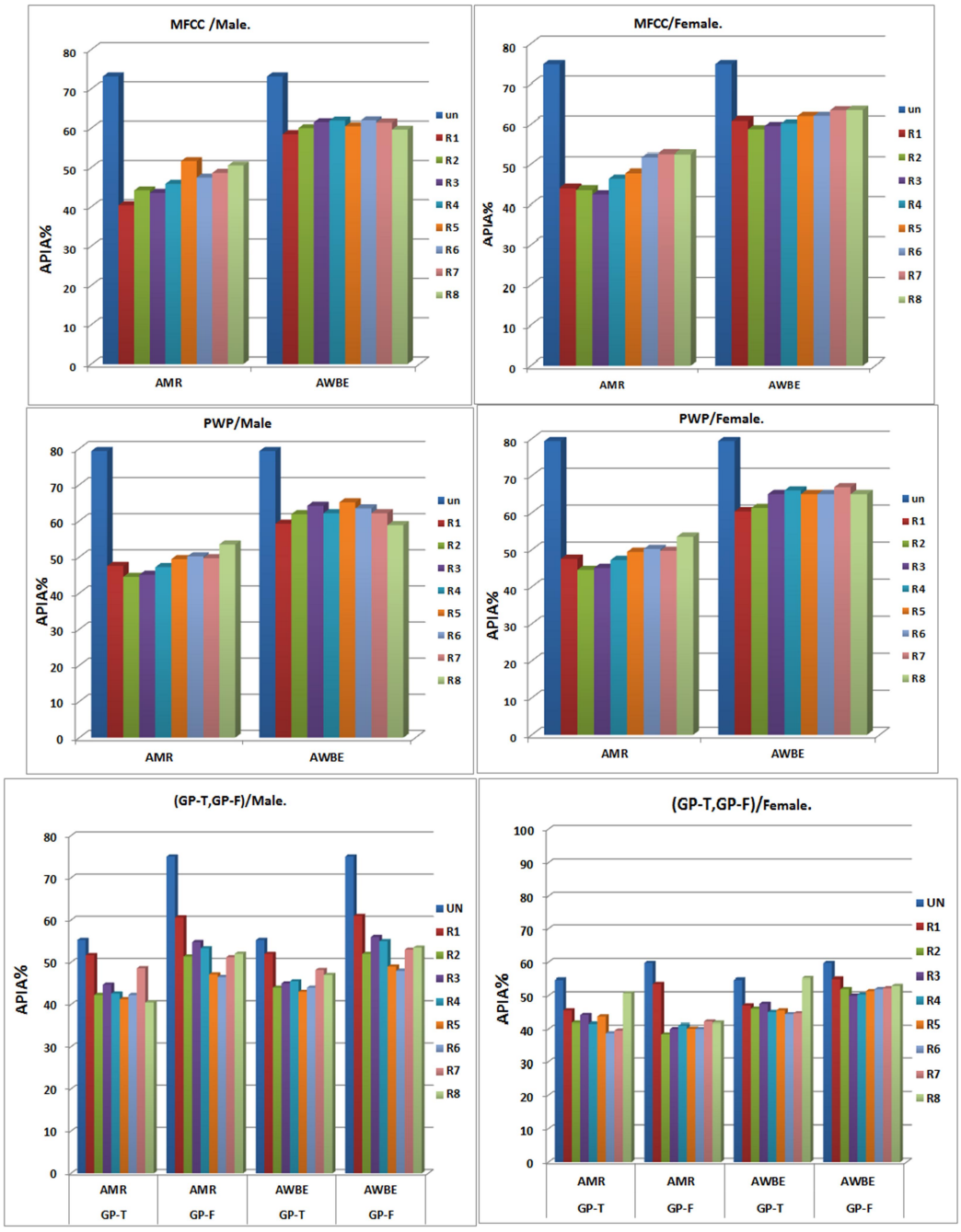

Figure 8.3: Achieved APIA values for uncompressed full-band speech (Un), compressed speech with eight different narrow-band AMR compression rates R1–R8, and for the extended-bandwidth speech (ABE) generated from AMR compressed speech with the same eight different bit rates R1–R8

8.5 Observations Based on Graphs Shown in Figure 8.3

8.5.1 Effect of ABE on AER Results for Different Speech Features

When using the MFCC features, an average increase in classification accuracy of 10% was observed due to the ABE for both male and female speakers across all compression rates R1–R8. For TEO-PWP features, the increase was around 14%–10% depending on gender and compression rates.

The glottal time frequency features showed 5% improvement due to the ABE, whereas the glottal frequency domain features improved their performance by 10%. In addition to showing the highest improvement from ABE, the TEO-PWP features provided the best overall performance in AER, which was consistent with previously reported results [38].

8.5.2 Effect of ABE on AER Results for Different Genders

For both genders, ABE led to an increase of AER accuracy. There were no significant differences between genders in emotion classification based on MFCCs and glottal time domain parameters.

For TEO-PWP features, the increase was around 14% for females and 10% for males. In contrast, the glottal frequency domain parameters were found to be slightly more effective with male voices than with female voices.

8.5.3 Effect of ABE on AER Results for Different AMR Compression Rates

For the MFCC features, both AMR and ABE AER showed monotonically increasing classification accuracy with increasing bit rates (40%–51% for AMR and 60%–64% for ABE). Since the first 12 MFCC parameters (indicative of the vocal tract spectral characteristics) were used, these results indicate that, at low bit rates, a lot of paralinguistic information embedded into the vocal tract spectrum was removed due to compression. Higher bit rates appeared to preserve more of this important information.

A similar increase was observed in the case of TEO-PWP features for male and female voices with AMR. The female ABE results showed almost constant performance for all bit rates ranging from R1 (4.75 kbps) to R8 (12.2 kbps). This indicates that all

compression rates appear to remove, to some extent, the spectral energy characteristics of emotional speech.

However, for the glottal time domain features, in both cases (AMR and ABE), there was a slow decrease in classification accuracy from R1 (4.75 kbps) to R6 (7.95 kbps), followed by a sudden increase in the level close to uncompressed speech. This increase was observed for R8 12.2 kbps in female voices and R7 (10.2 kbps) in male voices. This observation indicates that glottal time domain information preserved at high bit rates could be essential to correctly identify emotional speech.

In the case of glottal frequency domain features, there was only a small decrease in performance at R1 (4.75 kbps) compared to uncompressed speech. At other rates, performance decreased but remained almost constant for all rates (R2=5.15 kbps to R8=12.2 kbps). This shows that AMR compression appears to preserve emotional glottal frequency characteristics at very low bit rates.

The irregularity of effects caused by the glottal features could also result from the fact that AMR compression replaces the original glottal wave with an LP estimate. The estimation process could cause significant distortion to emotionally important features.

8.5.4 AER from Compressed and Uncompressed Speech

Although ABE improved the average accuracy of AER in all cases, the results were, on average, around 10% below the accuracy achieved for uncompressed speech. This indicates that the methodology used to achieve ABE was not able to fully reconstruct the emotional content of the uncompressed speech. This may be due to inadequacies of the ABE method and the loss of information from the reduction in bit rate from AMR. Therefore, further research is needed to improve the emotion-preserving aspect of ABE.

8.6 Conclusion

An Artificial Bandwidth Extension (ABE) method has been applied to test if improvement of emotion recognition from compressed narrow-band speech signal can be achieved. The original speech bandwidth was reduced using the AMR speech compression method.

The band extension was performed at the high frequency end of the narrow-band speech spectrum using spectral folding and spectral envelope estimation methods.

A standard benchmark approach was used to compare the accuracy of speech emotion recognition (AER) between uncompressed speech (UN), narrow-band compressed speech (AMR) and the speech with artificially extended bandwidth (ABE).

The AER results showed that application of the ABE lead to at least 5% improvement in emotion recognition accuracy compared to narrow-band compressed speech.

The amount of AER varied across different types of feature parameters, genders and compression rates. The MFCC and TEO parameters showed monotonic improvement of AER accuracy with increasing bit rates indicating that, the lower is the compression rate, the more important emotional information characterizing spectral characteristics of the vocal tract is removed from speech signal.

However, the glottal features didn't follow this pattern indicating that both, the glottal time domain information preserved at high bit rates and the glottal frequency domain information preserved at low bit rates are important for AER.

In all cases, the performance of the AER for ABE speech was below the recognition rates achieved with uncompressed speech. Therefore, further research is needed to improve the accuracy of the artificial bandwidth extension methodology .The ABE method was applied to test whether emotion recognition from compressed narrow-band speech signals can be improved. The original speech bandwidth was reduced using the AMR speech compression method.

Chapter 9: Discussion and Summary of Findings

9.1 Summary

This chapter summarises and discusses the major findings of this study. Suggestions of future research directions are given.

9.2 Discussion and Major Findings

9.2.1 What is the Effect of Speech Compression Techniques on AER?

It was shown that standard speech compression methods have a significant effect on the automatic emotion recognition (AER), and in general lead to degradation of AER accuracy.

The experiments included three different types of standard speech compression techniques (AMR, AMR-WB, AMR-WB+ and MP3) and three types of acoustic speech parameters (MFCC, TEO-PWP and GP-T&GP-F).

The modelling and classification of emotional speech was achieved using the GMM algorithm.

AMR-WB was found to lead to smallest degradation of AER accuracy followed AMR-WB+ and MP3, the worse performance was observed for the narrow band AMR.

These observations indicated the importance of high frequency speech components to AER.

As expected, lower bit rates which imply higher distortion to the speech signal; lead to lower AER accuracy indicating that low bit rates remove vital emotional cues from speech signals. These cues are most likely to be located at high frequency end of speech spectrum.

In contrast, codecs with higher bit rates which are likely to introduce less distortion provided higher accuracy of AER.

Although in general, the amount of degradation increased with the decreasing bit rates, there were some exceptions.

In particular, the combination of the AMR-WB 6.6kbps compression and the TEO-PWP features provided lead to results that were higher than for uncompressed speech. This outcome could be attributed to the fact that in this particular case, the compression patterns lead to an optimal data selection for AER. This could be similar to routinely used in AER data selection or data reduction techniques which usually lead to performance improvement.

The dependency patterns between the bit rates and emotion classification accuracy varied significantly across different genders, coding techniques and types of acoustic speech parameters used to distinguish between different emotions.

Generally, the classification results for all codecs and features, and across all bit rates, did not fall below 40%, which was significantly higher than the guessing threshold of 15% for the simultaneous recognition of seven classes of emotional speech.

One of the reasons for the observed degradation of emotional content in compressed speech could be the fact that current speech compression methods are optimised for maximum speech intelligibility. Therefore, no objectives are used to ensure that the paralinguistic (emotional) content is preserved and fully conveyed to listeners.

Described here results are consistent with previously reported effects of speech compression on speaker recognition [3], [4].

In particular [3] reported that narrow band AMR coding lead to a significant impact on speech parameters as a function of bit rate, however there was not a consistent trend.

Clear gender differences were observed due to AMR compression, likely caused by differences in pitch, with higher pitch female speech being affected significantly more by the codec than that of lower pitch male speech.

Coding systems (CELP, LPC and GSM) degraded significantly the perceptual speech quality of speech (formant and F0 trajectories) and subsequently the outcomes of speaker recognition [4].

It is therefore likely that observed here changes in AER could be caused by similar mechanisms.

9.2.2 What is the Effect of Band Reduction and Coding on AER?

It was investigated how a band limitation affects the accuracy of AER. The effects of these factors on AER were analysed using a range of different features that have recently been reported to provide high performance in speech emotion recognition. These features include MFCCs, TEO-PWP and GP-T&GP-F parameters. Acoustic class models were trained and classified using the GMM technique.

The results indicated that the low-frequency components (0 kHz to 1 kHz) of speech containing the fundamental frequency information, as well as the high-frequency components (above 4 kHz) play an important role in AER.

9.2.3 What is the Effect of Speech Compression and Hearing Loss Simulation on AER?

It was shown that band reduction of speech signal that simulates a typical high frequency hearing loss has a significant impact on AER and reduces AER accuracy.

These outcomes are consistent with findings described in 9.2.1 and 9.2.2, where removal of high frequency components from speech signal was shown to significantly reduce AER accuracy.

Given that, a typical age-related hearing loss is characterised by reduced ability to hear high frequency components of speech (see Figure 5.2), hearing aids users may not be able to capture full emotional contents of speech and subsequently experience reduced ability to recognise emotions from speech signals.

This effect may be even larger when an impaired hearing person listens to compressed speech.

Outcomes of these experiments indicated that this hypothesis could be true.

This experiment investigated effects of standard speech compression techniques on AER from speech modified in a way that simulates a typical hearing loss.

Significant reduction of AER accuracy was observed for uncompressed speech modified in a way simulating a typical mild-to-moderate high frequency hearing loss. This accuracy was further reduced when the modified speech was compressed.

It was shown that, for all types of feature parameters, compression methods and gender, the speech compression reduced the AER accuracy by about 10%-30% (depending on types of features and compression method) compared to uncompressed speech, and the subsequently applied hearing loss simulation decreased this accuracy even further by about 5%-10% and made it flat across compression rates.

For uncompressed speech the AER accuracy reduction due to HLS was about 20%-30% depending on types of features and compression method.

Therefore, the degradation of AER accuracy due to compression was found to be significantly increased by simulation of hearing loss.

Generally no significant differences between genders in the above trends were observed.

It is important to note that, these observations are based on machine learning. The full investigation would have to include subjective listening tests which were beyond the scope of this thesis.

9.2.4 What are the Combined Effects of Speech Compression and Noise on AER?

Experimental observations have shown that addition of noise to either uncompressed or compressed speech reduce accuracy of AER.

It was shown that the best performing under noisy conditions features were MFCCs and the best performing speech compression algorithms was AMR-WB.

The amount of accuracy reduction depended on SNR, type of speech features, gender and compression method.

Gender differences were observed, with male voices generally outperforming female voices. These differences become more prominent as the SNR values decrease. While for the AMR-WB and AMR-WB+ compression these differences are very small, they become clearly noticeable when the narrow-band AMR compression is applied.

These observations were consistent with previous reports of strong gender dependency in speech emotion [7], [19], [24], [33], stress [8], [18], [19], [20], [53] and depression classification [20], [21], [22], [23].

Higher resilience to noise observed in male voices could be attributed to the fact that male voices show generally lower values of fundamental frequency F0, as well as formant and harmonic frequencies compared to female voices. This means that (1) addition of high frequency noise may have lesser effect on male speech than female speech; (2) band reduction introduced by AMR compression may have smaller effect on male voices than female voices.

For uncompressed speech, the AER accuracy was reduced depending on the SNR value and type of speech feature parameters. For all values of SNR, the best performing features were MFCCs followed by the TEO-PWP, and the worse performing glottal parameters GP-T&GPF.

The same order of features' performance was observed for AER from compressed speech. Here again the MFCCs provided the highest performance followed by TEO-PWP and the worse performing GP-T&GP-F. The same order was holding for all values of SNR values and for both genders.

For all three SNR values (15dB, 10dB and 5dB) AER provided the highest accuracy values for the AMR-WB algorithm. The AMR-WB was followed by slightly worse performing AMR-WB+, and the least performing narrow-band AMR.

This means that the AMR-WB was found to be the best performing and therefore, the most robust or noise resilient speech compression method for AER.

High performance of the AMR-WB compression in AER can once again attributed to the fact that the algorithm preserves high bandwidth of speech which in terns implies that the emotional cues are most likely to be conveyed through high frequency components of speech signals.

If the high part of speech spectrum is preserved, addition of noise leads much smaller degradation of AER accuracy compared to cases when the high frequency information is either completely (AMR) or partially removed (AMR-WB+).

9.2.5 Is it Possible to Mitigate Detrimental Effects of Speech Compression on AER Using Speech Spectrogram Features?

It was shown that, speech features representing full spectrogram (SS) show only very small (5%) degradation of AER accuracy when applied to speech compressed with AMR-WB and AMR-WB+.

The high performance of the SS features was consistently observed across all tested compression rates.

These results are generally consistent with previous reports describing state of the art performance of speech spectrograms [70]. However this study for the first time shows very robust performance of the spectrogram features under speech compression conditions.

Incorporation of human auditory perception characteristics in the form of time-frequency features representing an average energy of either Critical Bands (SS-CB) or Bark Bands (SS-Bark) was found ineffective. It did not lead to an improvement of AER from compressed speech. This could be attributed to the fact that the auditory perception criteria have been defined assuming that the signal has full bandwidth [73]. Compression techniques such as AMR in particular, reduce the bandwidth and thus remove a large amount of important auditory cues.

Both SS-CB and SS-Bark features were outperformed by the SS features representing full speech spectrograms indicating that spectral energy features alone are not sufficient for AER and have to be supported by richer information given by complete spectrograms.

The AMR-WB compression technique lead to the smallest degradation of the AER accuracy indicating that other techniques degraded the performance either due to severe bandwidth reduction (AMR) or optimization for music rather than speech (AMR-WB+ and MP3).

MP3 which is the most popular audio compression format, was found to be particularly bad in preserving emotional speech aspect leading to very big (30%-40%) reduction of AER accuracy compared to uncompressed speech.

One of the possible explanations of such low performance of MP3 is that the code applies compression by eliminating sounds that are likely to be ignored by human ear according to psychoacoustic criteria. Original signal spectrum is divided into 32 frequency bands and FFT-based analysis is used to estimate Auditory Masking Threshold to determine sounds in each band below the masking threshold (they will be hidden by louder sounds at close frequencies). The algorithm is also looking if the signal is fairly constant, or does it change? Are there any sharp transient sounds that need to be preserved and which might mask other transients just before or after? This information is then used during the compression to figure out which information can be safely discounted.

Given that the psychoacoustic criteria have been derived using either ideal sounds such as sinusoids or clean speech with neutral emotional contents, some of the criteria used by MP3 could eliminate transient or nonstationary sounds that play important role in conveying emotional aspect of speech[51], [52], [73].

Observed gender differences in most cases lead to higher AER accuracy for male voices than for female voices. This effect can possibly attributed to the fact that for male voices the fundamental frequency and formant frequencies are positioned at lower frequency ranges than for female voices. This means that speech compression methods which reduce high frequency contents of speech are more likely to remove vital high frequency harmonics of F0 as well as high frequency formants. As indicated in Chapter 4, these high frequency components are likely to play important role in conveying emotional aspect of speech.

9.2.6 Is it Possible to Reduce Detrimental Effects of Narrow-Band Speech Compression on AER Using Artificial Bandwidth Extension (ABE) of Speech Signals?

The results have shown that by extending the narrow-band of AMR-compressed speech an improvement of AER accuracy can be achieved.

An Artificial Bandwidth Extension (ABE) method has been applied to test if improvement of emotion recognition from compressed narrow-band speech signal can be achieved. The original speech bandwidth was reduced using the AMR speech compression method.

The band extension was performed at the high frequency end of the narrow-band speech spectrum using spectral folding and spectral envelope estimation methods.

A standard benchmark approach was used to compare the accuracy of speech emotion recognition (AER) between uncompressed speech (UN), narrow-band compressed speech (AMR) and the speech with artificially extended bandwidth (ABE).

The AER results showed that application of the ABE lead to at least 5% improvement in emotion recognition accuracy compared to narrow-band compressed speech.

The amount of AER varied across different types of feature parameters, genders and compression rates. The MFCC and TEO parameters showed monotonic improvement of AER accuracy with increasing bit rates indicating that, the lower is the compression rate, the more important emotional information characterizing spectral characteristics of the vocal tract is removed from speech signal.

However, the glottal features didn't follow this pattern indicating that both, the glottal time domain information preserved at high bit rates and the glottal frequency domain information preserved at low bit rates are important for AER.

In all cases, the performance of the AER for ABE speech was below the recognition rates achieved with uncompressed speech. Therefore, further research is needed to improve the accuracy of the artificial bandwidth extension methodology.

9.3 Future Work

Future studies are needed to cross validate described here results using different data bases. Given the fact that the data based used here was generated by actors, it is paramount to conduct future experiments on speech data representing natural emotions, different languages, dialects and cultural backgrounds of speakers.

Future investigations should include speech signals sampled at higher frequencies with bandwidth extending beyond 8 kHz.

Majority of existing compression methods have been developed to optimise for speech intelligibility. To the best of our knowledge no experiments have been conducted to test their "emotional intelligibility". These tests could conduct a systematic in-depth

analysis of coding techniques combined with subjective hearing tests designed to test their emotion recognition scores.

Therefore development or improvement of existing compression method to preserve the complete linguistic as well as, emotional contents is an important area for future research.

The thesis has flagged out a possibility that hearing aids users may not be able to capture full emotional contents of speech due to reduction of high frequency speech components. New hearing aid solutions could be investigated to compensate for this deficiency.

In general, effects of various auditory conditions limiting the speech spectrum, such as hearing impairment and the use of hearing aids and cochlear implants on the SER, should be investigated.

References

[1] Byrne C, and Foulkes P, The mobile phone effect on vowel formants, Speech, *Language and the Law*, vol. 11, no. 1, pp. 83–10, 2004.

[2] Guillemin BJ, and Watson CI, Impact of the GSM AMR Speech Codec on Formant Information, Proceedings of the 11th Australasian International Conference on Speech Science and Technology, Auckland, New Zealand, 6–8 December 2006.

[3] Besacier L, Grassi S, Dufaux A, Ansorge M, and Pellandini F, GSM speech coding and speaker recognition, Proceedings of IEEE ICASSP, Istanbul, Turkey, 5–9 June 2000.

[4] Phythian M, Ingram J, and Sridharan S, Effects of speech coding on text-dependent speaker recognition, Proceedings of IEEE Region Ten Conference (Tencon ’97), Brisbane, Australia, December 1997.

[5] McClelland E, Familial similarity in voices, BAAP Colloquium, Glasgow, Scotland, April 2000.

[6] Paeschke A, and Sendlmeier W, Prosodic characteristics of emotional speech: measurements of fundamental frequency movements, Proceedings of ISCA ITRW on Speech and Emotion. Belfast, 2000, pp. 75–80.

[7] He L, Lech M, and Allen NB, On the importance of glottal flow spectral energy for the recognition of emotions in speech, Interspeech, 2010.

[8] He L, Lech M, Maddage N, and Allen N, Stress detection using speech spectrograms and sigma-pi neuron units, iCBBE ICNC’09–FSKD’09, Tianjin, China, 14–16 August 2009.

[9] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, and Weiss B, A database of German emotional speech, Interspeech, 2005.

[10] Särndal, C-E et al., Stratified sampling: model assisted survey sampling, New York, NY, Springer, pp. 100–109, 2003.

[11] ETSI, Digital cellular telecommunications system (Phase 2+), adaptive multi-rate (AMR) speech transcoding. ETSI-EN-301-704 V7.2.1 (2000-04), 2000.

[12] ETSI TS 126.090 (V9.0.0 2010-01), adaptive multi-rate (AMR) speech codec, transcoding functions (3GPP TS26.090 version 9.0.0 Release 9).

[13] ETSI TS 126.190 (V9.0.0 2010-01), adaptive multi-rate wideband (AMR-WB) speech codec, transcoding functions (3GPP TS 26.190 version 9.0.0 Release 9).

[14] ETSI TS 126 304 V11.0.1 (2012-10), extended adaptive multi-rate wideband (AMR-WB+) codec, floating-point ANSI-C code (3GPP TS 26.304 version 11.0.1 Release 11).

[15] Grimm M, Kroschel K, Mower E, and Narayanan S, Primitives-based evaluation and estimation of emotions in speech, *Speech Communication*, vol. 49, no. 10–11, pp. 787–800, 2007.

[16] Krothaplli RS, and Koolagudi GS, Emotion recognition using speech features, New York, NY, Springer Science+Business Media, 2013.

[17] Mubarak OM, Ambikairajah E, and J Epps, Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources, in the 8th International Symposium on Signal Processing and its Applications, Sydney, Australia, 28–31 August 2005.

[18] Zhou G, Hansen JHL, and Kaiser JF, Nonlinear feature based classification of speech under stress, *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[19] He L, Stress and emotion recognition in natural speech in the work and family environments, Ph.D. thesis, Department of Electrical Engineering, RMIT University, Melbourne, November 2010.

[20] Lech M, Song I, Yellowlees P, and Diederich J (Eds), Mental health informatics, *Studies in Computational Intelligence*, vol. 491, Springer Verlag, 2014.

[21] Low LSA, Maddage NC, Lech M, Sheeber LB, and Allen NB 2011, Detection of clinical depression in adolescents' speech during family interactions, *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[22] Moore E, Clements MA, Peifer J, and Weisser L, Critical analysis of the impact of glottal features in the classification of clinical depression in speech,

IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 96–107, 2008.

[23] Ooi KEB, Lech M, and Allen NB, Multi-channel weighted speech classification system for prediction of major depression in adolescents, *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 2, pp. 497–506, 2013.

[24] He L, Lech M, Jing Zhang J, Ren X, and Deng L, Study of wavelet packet energy entropy for emotion classification in speech and glottal signals, Proceedings of Fifth International Conference on Digital Image Processing, 19 July 2013, doi:10.1117/12.2030929.

[25] Gelfand SA, Hearing: an introduction to psychological and physiological acoustics, 4th ed., New York, Marcel Dekker, 2004.

[26] Maragos P, Kaiser JF, and Quatieri TF, Energy separation in signal modulations with application to speech analysis, *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[27] Airas M, TKK aparat: an environment for voice inverse filtering and parameterization, *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[28] Ballabio D, and Consonni V, Classification tools in chemistry—part 1: linear models, *Analytical Methods*, vol. 5, pp. 3790–3798, 2013.

[29] Ballabio D, and Todeschini R, Multivariate classification for qualitative analysis, in *Infrared Spectroscopy for Food Quality Analysis and Control*, Elsevier, 2009.

[30] Iliev AI, Scordilis MS, Papa JP, and Falcão AX, Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*. In Press, 2009.

[31] Neiberg D, Elenius K, and Laskowski K, Emotion recognition in spontaneous speech using GMMs, in Interspeech, 2006, Pittsburgh, Pennsylvania, pp. 809–812, 17–19 September 2006.

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

[33] He L, Lech M, Maddage N, and Allen N, 'Emotion recognition in natural speech using empirical mode decomposition and renyi entropy, International Symposium on Bioelectronics and Bioinformatics IBBS'09, Melbourne, 9–11 December 2009.

[34] Yildirim S, Narayanan S, and Potamianos A, Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, vol. 25, no. 1, pp. 29–44, 2011.

[35] Lajevardi SM, and Lech M, Facial expression recognition using a bank of neural networks and logarithmic gabor filters, Conference Proceedings, DICTA 2008, Canberra, 1–3 December 2008.

[36] Lajevardi SM, and Lech M, Facial expression recognition from image sequences using optimised feature selection, Conference Proceedings, IVCNZ 2008, Christchurch, New Zealand, 26–28 November 2008.

[37] Lajevardi SM, and Lech M, Averaged gabor filter features for facial expression recognition, Conference Proceedings, DICTA 2008, Canberra, 1–3 December 2008.

[38] Albahri A, Lech M, and Cheng E, Effect of speech compression on the automatic recognition of emotions, *International Journal of Signal Processing Systems*, vol. 4, no. 1, pp. 55–61, 2016.

[39] Leslie S, and Greenberg JDS, Emotion in psychotherapy: affect, cognition, and the process of change, 1987.

[40] Thayer RE, The biopsychology of mood and arousal, New Yok, NY: Oxford University Press, 1989.

[41] Petrushin VA, Emotion recognition in speech signal: experimental study, development, and application, Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, 2000.

[42] Batliner A, Fischer K, Huber R, Spilker J, and Nöth E, How to find trouble in communication, *Speech Communication*, vol. 40, no. 1–2, pp. 117–143, 2003.

[43] Ramakrishnan, S, Recognition of emotion from speech: a review. Speech Enhancement, Modeling and recognition–algorithms and Applications, pp. 120–136, 2012.

[44] Laaksonen L, Artificial bandwidth extension of narrowband speech-enhanced speech quality and intelligibility in mobile devices, thesis, School of Electrical Engineering, Aalto University, May 2013.

[45] Laaksonen L, Pulakka H, Myllylä V, and Alku P, Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal, *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 780–787, 2009.

[46] Pulakka H, Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech, Ph.D. thesis, School of Electrical Engineering, Aalto University, 2013.

[47] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, and Taylor J, Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[48] Ververidis D, and Kotropoulos C, Emotional speech recognition: resources, features, and methods, *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[49] Fernandez R, A computational model for the automatic recognition of affect in speech, Ph.D. dissertation, School of Architecture and Planning, Massachusetts Institute of Technology, 2004.

[50] DSP System Toolbox, MathWorks (25 July 2016), available at http://au.mathworks.com/help/dsp/index.html

[51] Gonzalez J, and Cervera T, The effect of MPEG audio compression on a multi-dimensional set of voice parameters, *Logopedics Phoniatrics Vocology*, vol. 26, pp. 124–138, 2001.

[52] MP3 (MPEG Layer III Audio Encoding). Digitalpreservation.gov. 2012-03-02.

[53] He L, Lech M, Memon S, and Allen N, Detection of stress in speech using perceptual wavelet packet analysis, GESTS International Transactions on Computer Science and Engineering, SUNJIN Publishing Co., Euljiro, Seoul, Korea, vol. 45, no. 1, pp. 17–24.

[54] Hidden Markov Toolkit, available at http://htk.eng.cam.ac.uk/.

[55] AngelHansen JHL, and Bou-Ghazale S, Getting started with SUSAS: a speech under simulated and actual stress database, Proceedings of Eurospeech-97, Rhodes, Greece, vol. 4, pp. 1743–1746, 1997.

[56] Schuller B, Müller R, Lang M, and Rigoll G, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles, Interspeech, 2005, ISCA, Lisbon, Portugal, pp. 805–809, 2005.

[57] DSP System Toolbox, MathWorks (11 August 2016), available at http://au.mathworks.com/help/comm/ref/awgn.html.

[58] Kleinschmidt M, Methods for capturing spectro-temporal modulations in automatic speech recognition, Acta Acustica, 2001.

[59] Chih T, Ru P, and Shamma S, Multiresolution spectrotemporal analysis of complex sounds, JASA, 2005, 118, pp. 887–906.

[60] Bouvrie J, Ezzat T, and Poggio T, Localized spectro-temporal cepstral analysis of speech, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[61] Ezzat T, Bouvrie J, and Poggio T, Spectro-temporal analysis of speech using 2-D Gabor filters, Interspeech, 2007.

[62] Ezzat T, and Tomaso PT, Discriminative word-spotting using ordered spectro-temporal patch features, Interspeech, 2008.

[63] Ezzat T, Bouvrie J, and Poggio T, AM-FM demodulation of spectrograms using localized 2D max-Gabor analysis. ICASSP, Hawaii, 2007.

[64] Meyer B, Kleinschmidt M, Robust speech recognition based on localized spectro-temporal features, ESSV, Karlsruhe, 2003.

[65] Tsang-Long P, and Wen-Yuan L, Comparison of several classifiers for emotion recognition from noisy Mandarin speech, Intelligent Information Hiding and Multimedia Signal Processing, 2007.

[66] Smith JO, and Abel JS, Bark and ERB bilinear transforms, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[67] RFC 4867—RTP payload format and file storage format for the adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-WB) audio codecs, p. 35.

[68] Quatieri TF, Discrete-time speech signal processing: principles and practice, Prentice Hall, 2001.

[69] Pierre-Yves O, The production and recognition of emotions in speech: features and algorithms, *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[70] Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. Multimedia, IEEE Transactions on, 16, 2203-2213.

[71] Osgood C., Suci G., Tannenbaum P., The Measurement of Meaning. Urbana, IL: Univ. Illinois Press., 1957.

[72] Kleinschmidt M., Hohmann V., Sub-band SNR estimation using auditory feature processing. Speech Communication, 2003. 39(1-2): p. 47-63.

[73] Moore B.C.J., Glasberg B.R., Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. , 1983. 74: p. 750-753.

[74] Matlab documentation:http://au.mathworks.com/help/comm/ref/awgn.html

[75] AngelSim^{TM}(TigerCIS): Cochlear Implant and Hearing Loss Simmulator, available from: http://www.tigerspeech.com/angelsim/angelsim_about.html

[76] Dillon H, Hearing Aids, Boomerang Press, Second Edition, 2010.

[77] Facts on Hearing Loss: http://hearnet.org.au/hearing-loss/facts-on-hearing-loss

[78] Stevens, Stanley Smith; Volkmann; John & Newman, Edwin B. (1937). "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America 8 (3): 185–190.

[79] Zhao X, Ahang S, Lei B., Robust emotion recognition in noisy speech via sparse representation, Neural Computing and Applications, June 2014, Volume 24, Issue 7, pp 1539–1553.

[80] Zhao X, Zhang S, Spoken emotion recognition via locality-constrained kernel sparse representation, Neural Computing and Applications, (2015) 26:735–744

[81] H.M. Fayek, M. Lech and L. Cavedon, “Towards real-time Speech Emotion Recognition using deep neural networks," in Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on, pp.1-5, 14-16 Dec. 2015.

[82] Busso C, Sungbok L, Narayanan S (2009) Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans Audio Speech Lang Process 17(4):582–596.

[83] Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun 49(3):201–212.

[84] Schuller B, Arsic D, Wallhoff F, Rigoll G (2006) Emotion recognition in the noise applying large acoustic feature sets. In: Speech Prosody, Dresden, Germany

[85] You M, Chen C, Bu J, Liu J, Tao J (2006) Emotion recognition from noisy speech. In: IEEE international conference on multimedia and expo (ICME’06), Toronto, Ont, pp 1653–1656

[86] SongM, YouM, LiN,ChenC(2008)Arobustmultimodal approach for emotion recognition. Neurocomputing 71(10–12):1913–1920

[87] Yeh L, Chi T (2010) Spectro-temporal modulations for robust speech emotion recognition. In: INTERSPEECH-2010, Makuhari, Chiba, Japan, pp 789–792

[88] Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory (4):1289–1306

[89] Baraniuk RG (2007) Compressive sensing [lecture notes]. IEEE Signal Process Mag 24(4):118–121

[90] Candes EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25(2):21–30

[91] Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227

[92] Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S (2010) Sparse representation for computer vision and pattern recognition. Proc IEEE 98(6):1031–1044 1552 Neural Comput & Applic (2014) 24:1539–1553

[93] Wagner A, Wright J, Ganesh A, Zhou Z, Mobahi H, Ma Y (2011) Towards a practical face recognition system: robust alignment and illumination by sparse representation. IEEE Trans Pattern Anal Mach Intell 99:1–15

[94] Sainath TN, Ramabhadran B, Nahamoo D, Kanevsky D, Sethy A (2010) Sparse representation features for speech recognition. In: INTERSPEECH-2010, Makuhari, Chiba, Japan, pp 2254–2257

[95] Gemmeke J, Virtanen T, Hurmalainen A (2011) Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans Audio Speech Lang Process 19(7):2067–2080.

[96] Souza PE, Effects of Compression on Speech Acoustics, Intelligibility, and Sound Quality, Trends in Amplification, SEGE publishers 2002, online version: http://tia.sagepub.com/content/6/4/131

[97] Jovicic ST, Jovanović N, Subotić M, Groz D, Impact of mobile phone usage on speech spectral features: some preliminary findings, International Journal Of Speech Language and the Law, VOL 22, NO 1 (2015)

[98] Künzel, H. J. (2001) Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics 8(1): 80–99.

[99] P. Petta, C. Pelachaud, and R. Cowie, Emotion-Oriented Systems: Springer-Verlag Berlin Heidelberg, 2011.

[100] F. Eyben, M. Wollmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," Journal on Multimodal User Interfaces, vol. 3, no. 1-2, pp. 7{19, 2010.

[101] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, The geneva

minimalistic acoustic parameter set (gemaps) for voice research and affective computing," Aective Computing, IEEE Transactions on, vol. PP, no. 99, pp. 1-10, 2015.

[102] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, May 2011, pp. 5688-5691.

[103] Li, Longfei; Zhao, Yong; Jiang, Dongmei; Zhang, Yanning; Wang, Fengna; Gonzalez, Isabel; Enescu, Valentin; Sahli, Hichem, Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013). IEEE, 2013. p. 312-317.

[104] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F.Wang, I. Gonzalez, E. Valentin, and H. Sahli, Hybrid deep neural network-hidden markov model (dnn-hmm) based speech emotion recognition, in A ective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, Sept 2013, pp. 312-317.

[105] C. Busso, S. Lee, and S. S. Narayanan, Using neutral speech models for emotional speech analysis." in Interspeech, 2007, pp. 2225-2228.

[106] K. Han, D. Yu, and I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech 2014, September 2014.

[107] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks," Multimedia, IEEE Transactions on, vol. 16, no. 8, pp. 2203-2213, Dec 2014.

[108] Stolar MN, Lech M and Burnett I, “Optimized multi-channel deep neural network with 2D graphical representation of acoustic speech features for emotion recognition”, ICSPCS-2014, Brisbane, Australia, pp. 1-6.

[109] L. Longfei, Z. Yong, J. Dongmei, Z. Yanning, W. Fengna, I. Gonzalez, et al., "Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition," in Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, 2013, pp. 312-317.

[110] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504-7, Jul 28 2006.

[111] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, ed: Curran Associates, Inc., 2012, pp. 1097-1105.

[112] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 05/28/print 2015.

[113] G. Hinton, D. Li, Y. Dong, G. E. Dahl, A. Mohamed, N. Jaitly, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," Signal Processing Magazine, IEEE, vol. 29, pp. 82-97, 2012.

[114] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, pp. 14-22, 2012.

[115] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet et al., The interspeech 2012 speaker trait challenge." in INTERSPEECH, 2012.

[116] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Muller, and S. S.Narayanan, The Interspeech 2010 paralinguistic challenge." in INTERSPEECH, 2010, pp. 2794-2797.

[117] M. Tahon and L. Devillers, Towards a small set of robust acoustic features for emotion recognition: Challenges," Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol. 24, no. 1, pp. 16-28, 2016.

[118] H. Fayek, “Literature Review”, Unpublished Internal Report, RMIT, 2015.

[119] Vlasenko B., Schuller B , Wendemuth A., Rigol G., “Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing”, Elsevier, Affective Computing and Intelligent Interaction, Volume 4738 of the series Lecture Notes in Computer Science pp 139-147.

[120] P. Kuppens, L.B. Sheeber, M.B. Yap, S. Whittle, J.G. Simmons, and N.B. Allen, “Emotional Inertia Prospectively Predicts the Onset of Depressive Disorder in Adolescents,” Emotion, vol. 12, no. 2, pp. 283-289, 2012.

[121] P. Kuppens, N.B. Allen, and L.B. Sheeber, “Emotional Inertia and Psychological Maladjustment,” Psychological Science, vol. 21, no. 7, pp. 984-991, 2010.

[122] C. Sandoval Rodriguez, “Evaluation of Artificial Bandwidth Extension Techniques for Mobile Telephone Speech”, EEET2312 Research Project, MC233 Master of Engineering (Electronic Engineering), RMIT June 2015.

[123] The current state of the art in this methodology belongs to recently published Microsoft Emotion API https://www.microsoft.com/cognitive-services/enus/emotion-api"