Effect of Speech Compression on the Automatic Recognition of Emotions

A. Albahri, M. Lech, and E. Cheng
School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia
Email: s3308189@student.rmit.edu.au, {margaret.lech, eva.cheng}@rmit.edu.au

***Abstract*—This paper investigates the effects of standard speech compression techniques on the accuracy of automatic emotion recognition. Effects of Adaptive Multi-Rates (AMR), Adaptive Multi-Rate Wideband (AMR-WB) and Extended Adaptive Multi-Rate Wideband (AMR-WB+) speech codecs were compared against emotion recognition from uncompressed speech. The recognition methods included techniques based on three different types of acoustic speech parameters: Teage Energy Operator features (TEO), Mel Frequency Cepstral Coefficients (MFCCs), and Glottal Time and Frequency domain features (GP-T&GP-F). The results showed that in general, all three speech compression techniques resulted in the reduction of emotion recognition accuracy. However, the amount of degradation varied across compression methods and types of acoustic features. It was observed that the accuracy of emotion recognition using the AMR-WB technique was higher than the accuracy of the AMR-WB+ and the AMR codecs. Further, the TEO-PWP features showed much more robust performance under different compression rates than the MFCC, GP-T and GP-F features.**

***Index Terms*—speech compression, emotion recognition, speech classification**

I. Introduction

The automatic recognition of emotions in speech has many applications in various human-machine communication systems, speaker recognition and verification, biometric security purposes, as well as medical and physiological services. However, 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 automatic emotion recognition rates have not yet been addressed.

Speech compression introduces many industrial advantages for telecommunications and speech technology, which support and serve speech recognition for human-to-machine communication. Such advantages include the reduction in delay for data transmission using telephony, the 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.

Some of the possible factors associated with speech compression that could affect the automatic emotion recognition include spectral modifications to speech signal introduced during the coding and decoding procedures. Another important factor is the limited bandwidth that is used by some coding techniques. These factors can alter dramatically the acoustic speech characteristics and impact directly on the accuracy of emotion recognition in speech. In [4], the Code-Excited Linear Prediction (CELP) and Linear Prediction (LP) based GSM speech codecs have been 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 of the F0 value by up to 30Hz making it closer to the F0 extracted from a landline uncompressed speech [5]. Furthermore, the F1 formants of vowels extracted form compressed speech were higher than F1 formants 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 Adaptive Multi Rate (AMR) speech codec. Interestingly, not all acoustic speech features perform worse with compressed speech. For example, the speech recognition accuracy has been shown to be improved when using speech features such as the Mel Frequency Cepstral Coefficients (MFCC) estimated from speech compressed by the GSM speech codec in comparison with the uncompressed speech [3], [6]. However, some of the limitations of these studies were the use of only the narrowband GSM AMR speech codecs (300-4300Hz), and a focus on only the classical speech features in the analysis of effects of speech compression. Despite the recent interest in automatic emotion recognition research, there are no comprehensive studies investigating the effect of speech compression on the affective characteristics of speech.

This study aims to address this gap and investigate how the standard speech compression techniques impact the accuracy of automatic emotion recognition. The current study extends the previous investigations into the effects of coding methods based not only on the narrow band AMR but also on the wideband AMR-WB and

Manuscript received August 21, 2014; revised November 21, 2014.

doi: 10.12720/ijsps.4.1.55-61

extended wideband AMR-WB+ speech codecs. The effects of these codecs are analyzed using a range of different features, recently reported to provide high performance in speech emotion recognition [7], [8]. These features include the Teager Energy Operator parameters (TEO-PWP), the Mel Frequency Cepstral Coefficients (MFCC), the glottal time parameters (GP-T) and the glottal frequency parameters (GP-F).

II. Method

A. Speech Database

The emotion recognition experiments were conducted on the Berlin Emotional Speech (BES) database described in [9]. The database contains speech samples representing 7 categorical emotions (anger, happiness, sadness, fear, disgust, boredom and neutral speech) spoken by 10 professional actors (5 female and 5 male) fluent in German. Each speaker simulated all 7 emotions while pronouncing 10 different utterances 5 short (2-4 seconds) and 5 long (5-9 seconds)), with linguistically neutral contents. The sampling frequency of the speech samples was 8kHz. Table I provides the numbers of available speech samples for different emotions.

TABLE I. Description of the Speech Data

	Ang	Bor	Disg	Fear	Hap	Neu	Sad
Male	60	34	8	26	21	38	17
Female	67	45	30	29	37	40	36

B. Experimental Framework

The speech samples representing either compressed or uncompressed speech were normalized into the range ±1. After removal of noise, and voiced/silence detection, the voiced speech frames were concatenated and used in the two-stage processing illustrated in Fig. 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.

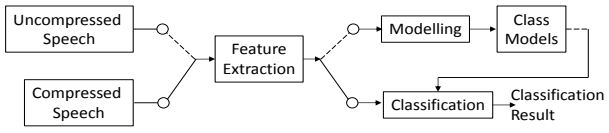

Figure 1. Block diagram of the experimental framework.

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% percent of data was used in the training process and 20% used in the testing. The classification results were assessed using the Average Percentage of Identification Accuracy (APIA) given in (1) [10]:

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

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. The emotion recognition was tested for each gender separately and Table II shows the compression bit rates tested in the experiments. Note that the compression rates in Table II 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 III.

TABLE II. Bit-Rates Used in the Experiments for Different Speech Compression Systems

Codec	Bit Rates (kbit/second)							
	R1	R2	R3	R4	R5	R6	R7	R8
AMR	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
AMR-WB	6.6	8.85	12.65	14.25	15.85	18.25	19.85	23.85
AMR-WB+	10.4	12	13.6	15.2	16.2	19.2	20.8	24

C. Speech Compression Methods

Adaptive multi rate (AMR) narrowband speech codec [11], [12]: AMR is based on Algebraic Code Excited Linear Prediction (ACELP), and has 8 narrow band modes (ranging from 300 to 3400KHz). Each of the 8 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) 12.2kbit/s. The speech is coded frame-by-frame with a frame size of 20ms (160 speech samples at 8kHz sampling rate). For each speech frame, the speech signal is analyzed using Linear Prediction (LP) of order 10 to calculate the LP coefficients, the adaptive codebook and the 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.

Adaptive multi rate wideband (AMR-WB) codec [13]: AMR-WB is an extension of AMR, with the wideband range of (50-7KHz) and 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 or 23.85kbit/s. Like AMR, AMR-WB is based on the ACELP coding technique. However, AMR-WB uses a 16th order LP short-term prediction filter and for each frame, the LP parameters, adaptive and fixed codebooks and the gain are calculated. These parameters are encoded and transmitted as the speech frame divides into sub-frames. The adaptive and fixed codebook parameters are transmitted for every sub-frame.

Adaptive multi rate wideband extension (AMR-WB+) codec [14]: The AMR-WB+ extends the AMR-WB method by adding transform coded excitation (TCX), bandwidth extension and stereo. While AMR and AMR-WB are optimized 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 12800Hz to 38400Hz. 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 (10.4Kbit/s), AMR-WB+ 240 bit/frame (12.0kbitit/s), AMR-WB+ 272 bit/frame (13.6kbitit/s), AMR-WB+ 304 bit/frame (15.2kbitit/s), AMR-WB+ 336 bit/frame (16.8kbit/s), AMR-WB+ 384 bit/frame (19.2kbit/s), AMR-WB+ 416 bit/frame (20.8kbit/s) and AMR-WB+ 480 bit/frame (24kbit/s). The current study applied only the mono rates due to the nature of the Berlin Emotional Speech database containing only mono recordings [9]

D. 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 paragraphs explain the feature extraction techniques applied to both compressed and uncompressed speech.

Mel frequency cepstral coefficients (MFCCs): The Mel Frequency Cepstral Coefficients (MFCCs) are some of the most frequently used features shown to provide good performance in speaker recognition and emotion classification in speech [15]-[17]. For each frame, the Fourier transform and the energy spectrum were estimated and mapped onto the mel-frequency scale. The Discrete Cosine Transform (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.

Teager energy operator features (TEO-PWP): Features derived from the Teager Energy Operator (TEO) [18] have been previously applied in emotion [7], stress [19], [20] and depression [21]-[24] classification systems. The process of calculating the TEO parameters followed the frame-based method introduced in [25], which calculates the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through the Perceptual Wavelet Packet (PWP) analysis as close estimates of the critical bands characterising the human auditory system [26]. For each frame of length 256 samples, values of the TEO instantaneous energy of a given signal x[n] were calculated using (2) proposed by Kaiser [27].

$$\Psi(x[n]) = x^2[n] - x[n+1]x[n-1] \quad (2)$$

The instantaneous energy was then used to evaluate the TEO autocorrelation function values using (3) [20].

$$R_{\Psi(x)}[k] = \frac{1}{2M+1}\sum_{n=-M}^{M}\Psi(x[n])\Psi(x[n+k]) \quad (3)$$

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.

Glottal time and frequency domain features (GP-T&GP-F): Glottal features have been shown to provide efficient classification of emotion [7] and depression [22]-[24] in speech. An Iterative Adaptive Inverse Filtering algorithm (IAIF) based on the discrete all-pole modeling (DAP) was used to generate the glottal wave, and the glottal parameters were calculated using procedures included in the TTK Aparat Toolbox [28]. The glottal time domain features (GP-T) were represented by 9 different parameters describing amplitudes, timing and duration of the opening and closing phases of the vocal folds. The glottal frequency domain features (GP-F) included 3 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.

E. Modelling and Classification Methods

The modelling and classification tasks were achieved using the Gaussian Mixture Model (GMM) algorithm, which has been effectively used in speech modeling in various speech recognition tasks [29]-[32], [9], [33]. 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 expectation maximization (EM) algorithm was applied to estimate the optimal values of these parameters. The Gaussian mixture modeling (or training) stage was integrated with the Bayesian classification decision procedure which determined the most probable classes for given query samples. A 3rd order Gaussian mixture model combined with the EM algorithm and the Bayesian classifier from the HTK toolbox were implemented to test the automatic classification of 7 different emotional categories using compressed and uncompressed speech and different types of feature parameters.

III. Results and Discussion

The following sections show how the three different speech compression techniques (AMR, AMR-WB and AMR-WB+) 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 Fig. 2-Fig. 7 separately for male and female speakers.

A. Classification Outcomes for Uncompressed Speech

The emotion classification task aiming to distinguish simultaneously between 7 different emotional classes represented a significant challenge. The aim was to achieve results that do not fall below the pure guess level which in this case was about 15%. The classification results for the uncompressed speech (Fig. 2-Fig. 7) 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 lead to around 73% (Fig. 2 and Fig. 5), the TEO-PWP - 78% (Fig. 3 and Fig. 6) and the GP-T - 55% (Fig. 4 and Fig. 7) of the classification accuracy. Although, the TEO-PWP provided the best performance, there was also a good performance given by the MFCCs. The glottal frequency domain parameters GP-F outperformed the GP-T in both genders (Fig. 4 and Fig. 7). The GP-F features were

significantly more effective with male voices than with female voices. In particular the GP-F led to 75% accuracy for male voices (Fig. 4) and only 59% accuracy for female voices (Fig. 7). These results were consistent with previously reported emotion recognition outcomes based on the uncompressed speech [7], [34], [17], [35].

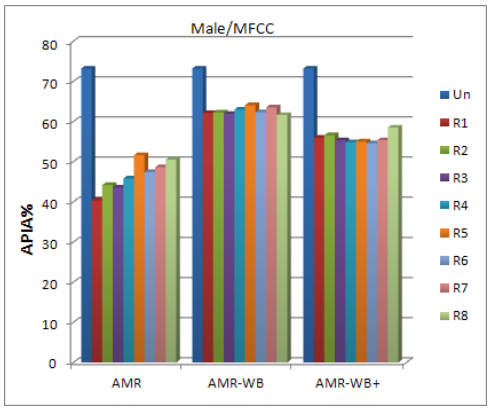

Figure 2. Average accuracy of multi-class emotion recognition for male speakers using MFCC features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

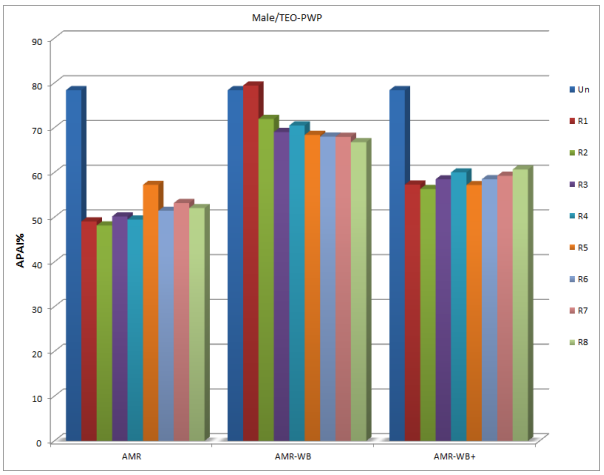

Figure 3. Average accuracy of multi-class emotion recognition for male speakers using TEO-PWP features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

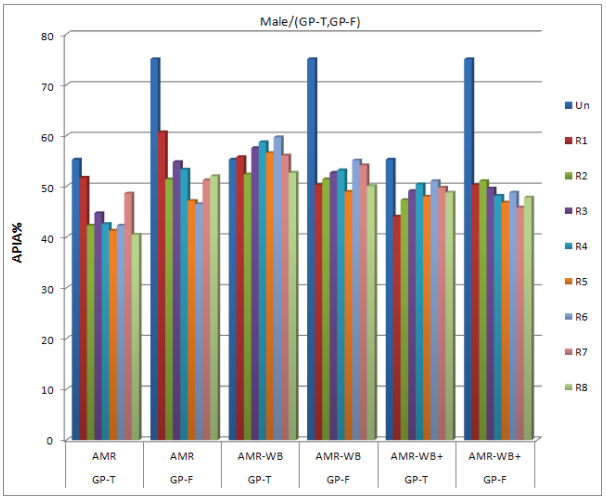

Figure 4. Average accuracy of multi-class emotion recognition for male speakers using GP-T&GP-F features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

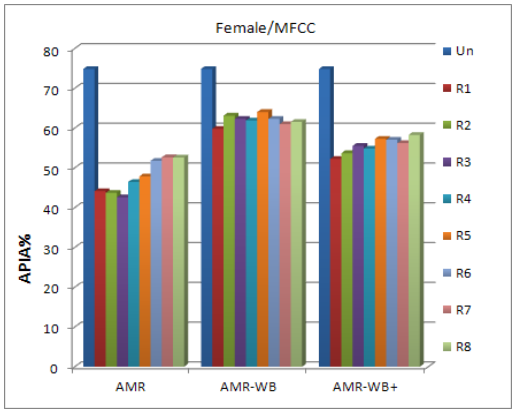

Figure 5. Average accuracy of multi-class emotion recognition for female speakers using MFCC features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

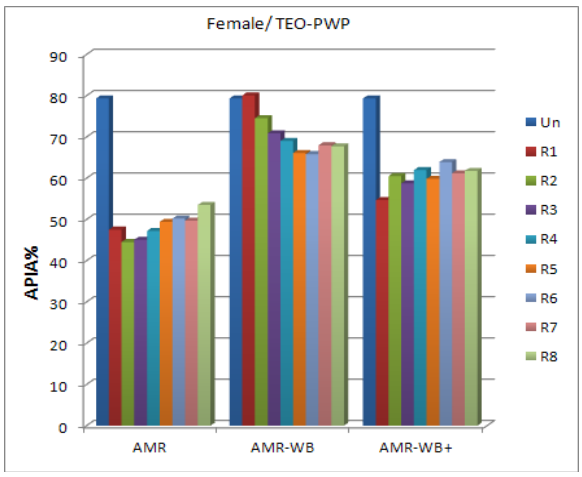

Figure 6. Average accuracy of multi-class emotion recognition for female speakers using TEO-PWP features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

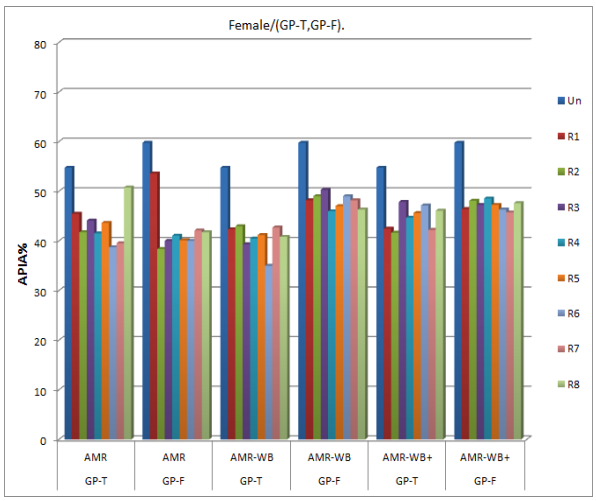

Figure 7. Average accuracy of multi-class emotion recognition for female speakers using GP-T&GP-F features; Un denotes uncompressed speech and R1-R8 are compression rates in an increasing order.

B. *Effect of the Narrow Band AMR Compression on Emotion Classification*

For the MFCCs, the AMR compression led to low classification accuracy 40%-51% (depending on the compression rate) compared to uncompressed speech.

There was a very clear decrease of the classification accuracy from 50% to 40% with the bit rates decreasing from R8 (12.2kbit/s) to R1 (4.75kbit/s). An outstanding 51% accuracy was observed for R5 (7.4kbit/s) in the case of male speech (see Fig. 2). Generally, there were no significant differences in these trends across genders.

For the TEO-PWP features, the classification accuracy dropped down to about 50% compared to the uncompressed speech (Fig. 3 and Fig. 6); however, the classification accuracy was almost the same (flat) for all bit rates from R8-R1. Like for AMR, an outstanding 57% accuracy was observed for R5 (7.4kbits/s) in the case of male speech (see Fig. 3). No other significant differences between genders were observed.

For the glottal features (GP-T&GP-F) in Fig. 4 and Fig. 7, the frequency parameters GP-F clearly outperformed the time domain parameters GP-T in both genders and the male voice classification achieved highest results than the female voice classification. Interestingly, the lowest bit rate R1 (4.75kbit/s) 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-R8 showed lower but almost flat performance compare 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 than for female voices.

C. *Effect of Wideband AMR-WB Compression on Emotion Classification*

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

For the TEO-PWP features, the classification accuracy was slightly increased for the lowest bit rate R1 (6.6kbit/s) to about 79% compared to the uncompressed speech (Fig. 3 and Fig. 6). An increase in the bit rate from R2 (8.85bit/s) to R8 (23.85bit/s) showed a clearly decreasing slope of the classification accuracy from about 74% (for R2) to 67% (for R8). There were no significant differences in these trends across genders.

The TEO-PWP results appear to contradict the informal belief that the lower are the compressed speech bit rates, the higher is the speech degradation and hence lower accuracy of emotion recognition. However, it is important to remember that the speech coding techniques used in this study were optimized for maximum speech intelligibility rather than for preserving the emotional contents. Moreover, previous studies of depression and emotion classification based on uncompressed speech indicated that the performance of the TEO features is highly dependent on the signal bandwidth [36], [23] and that the optimal feature selection, which is effectively a speech compression process, can lead to a significant improvement in emotion classification results [37]-[39]. A similar improvement over the uncompressed speech was also reported in [3], [6], where the 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 wide band condition associated with the AMR-WB 6.6 kbit/s 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 Fig. 4 and Fig. 7, showed a different performance for male and female speakers. For the male speaker, the time parameters GP-T slightly outperformed the frequency parameters (GP-F) with a consistent performance across all rates R1-R8 leading to 55% accuracy for the GP-T and 52% for the GP-F. In contrast, for the female speaker, the frequency parameters GP-F outperformed the time parameters (GP-T) with again almost flat performance across all rates R1-R8 leading to 40% accuracy for the GP-T and 48% for the GP-F.

D. *Effect of Extended Wide Band AMR-WB+ Compression on Emotion Classification*

For the MFCC parameters, The AMR-WB+ codec showed performance trends similar to the AMR-WB (Fig. 2 and Fig. 5). In all cases, the classification accuracy was slightly higher than for the AMR but lower than for the AMR-WB.

For the TEO-PWP and the AMR-WB+ in mono mode, the performance was slightly higher but in all trends similar to the AMR (Fig. 3 and Fig. 6), with classification accuracy slightly increasing with the increasing bit rate from R1 (10.4 bit/s) to R8 (24 bit/s).

The glottal features derived from the AMR-WB+ compression for both GP-T and GP-F exhibited very similar performance with almost flat accuracy(about 45% on average) across all rates R1-R8 (Fig. 4 and Fig. 7). There were no significant differences between genders.

IV. Conclusion

The current study investigated the effect of speech compression on the automatic simultaneous recognition of 7 types of emotional speech samples obtained from the Berlin Emotional Speech database. The experiments included three different types of standard speech compression techniques (AMR, AMR-WB and AMR-WB+) 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 intuitively predictable that lower bit rates imply higher distortion to the speech signal and as such are expected to remove some information about speech emotions and lead to lower accuracy of automatic emotion recognition. In contrast, codecs with higher bit rates introduce less distortion and therefore could be expected to provide higher accuracy of automatic emotion recognition.

The experimental results presented in this paper confirmed this general expectation, showing that speech compression based on standard codecs degrades the automatic emotion recognition outcomes. However, the amount of this degradation was not always increasing with the decreasing bit rates.

In particular, it was shown that the combination of the wide band AMR-WB 6.6kbit/s compression and the TEO-PWP features provide an optimal configuration for high accuracy multiclass emotion recognition in speech and led to results that were higher than for the uncompressed speech.

The dependency patterns between the bit rates and the 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, 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 7 classes of emotional speech.

One of the reasons for the observed degradation of emotional contents in compressed speech could be the fact that, the current speech compression methods are optimized for maximum speech intelligibility. Therefore, no objectives are used to ensure that the paralinguistic (emotional) contents are preserved and fully conveyed to the listeners. Future studies improving this aspect of speech coding standards are needed.

REFERENCES

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

[2] B. J. Guillemin and C. I. Watson, "Impact of the GSM AMR speech codec on formant information," in *Proc. 11th Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand, Dec. 6-8, 2006.

[3] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM speech coding and speaker recognition," in *Proc. IEEE ICASSP*, Istanbul, Turkey, Jun. 5-9, 2000.

[4] M. Phythian, J. Ingram, and S. Sridharan, "Effects of speech coding on text-dependent speaker recognition," in *Proc. IEEE Region Ten Conference (Tencon '97)*, Brisbane, Australia, Dec. 1997.

[5] E. McClelland, "Familial similarity in voices," in *Proc. BAAP Colloquium*, Glasgow, Scotland, Apr. 2000.

[6] A. Paeschke and W. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *Proc. ISCA ITRW on Speech and Emotion. Belfast*, 2000: pp. 75-80.

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

[8] L. He, M. Lech, N. Maddage, and N. Allen, "Stress detection using speech spectrograms and sigma-pi neuron units," in *Proc. ICBBE ICNC'09-FSKD'09*, Tianjin, China, Aug. 14-16, 2009.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, 2005.

[10] C.-E. Särndal, *et al.*, "Stratified sampling," in *Model Assisted Survey Sampling*, New York: Springer, 2003, pp. 100-109.

[11] ETSI (2000) Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding, ETSI-EN-301-704 V7.2.1, Apr. 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] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-Based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787-800, 2007.

[16] R. S. Krothaplli and G. S. Koolagudi, *Emotion Recognition Using Speech Features*, New York: Springer Science+Business Media, 2013.

[17] O. M. Mubarak, E. Ambikairajah, and J. Epps, "Analysis of an mfcc-based audio indexing system for efficient coding of multimedia sources," in *Proc. 8th International Symposium on Signal Processing and its Applications*, Sydney, Australia, Aug. 28-31, 2005.

[18] G. Zhou, J. H. L. Hansen, J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201-216, 2001.

[19] L. He, M. Lech, and N. B. Allen, "On the importance of glottal flow spectral energy for the recognition of emotions in speech," in *Proc. Interspeech*, 2010.

[20] L. He, "Stress and emotion recognition in natural speech in the work and family environments," Ph.D. thesis, Dept. Elec. Eng., RMIT University, Melbourne, Nov. 2010.

[21] M. Lech, I. Song, P. Yellowlees, and J. Diederich, "Mental health informatics," in *Studies in Computational Intelligence*, Springer Verlag, 2014.

[22] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574-586, 2011.

[23] E. Moore, M. A. Clements, J. Peifer, and L. Weisser, "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.

[24] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multi-Channel weighted speech classification system for prediction of major depression in adolescents," *IEEE Transactions on Biomedical Engineering*, Feb; vol. 60, no. 2, pp. 497-506, 2013.

[25] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng, "Study of wavelet packet energy entropy for emotion classification in speech and glottal signals," in *Proc. Fifth International Conference on Digital Image Processing*, 2013.

[26] S. A. Gelfand, *Hearing: An introduction to Psychological and Physiological Acoustics*, 4th ed., New York: Marcel Dekker, 2004.

[27] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024-3051, 1993.

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

[29] D. Ballabio and V. Consonni, "Classification tools in chemistry - Part 1: Linear models," *Analytical Methods*, vol. 5, pp. 3790-3798, 2013.

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

[31] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech and Language*, 2009.

[32] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. INTERSPEECH 2006 - ICSLP*, Pittsburgh, Pennsylvania, Sep. 17-19, 2006, pp. 809-812.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, 2005.

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

[35] L. He, M. Lech, N. Maddage, and N. Allen, "Emotion recognition in natural speech using empirical mode decomposition and renyi entropy," in *Proc. International Symposium on Bioelectronics and Bioinformatics IBBS'09*, Melbourne, Australia, Dec. 9-11, 2009.

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

[37] S. M. Lajevardi and M. Lech, "Facial expression recognition using a bank of neural networks and logarithmic gabor filters," in *Proc. DICTA*, Canberra, Australia, Dec. 1-3, 2008.
[38] S. M. Lajevardi and M. Lech, "Facial expression recognition from image sequences using optimised feature selection," in *Proc. IVCNZ*, Christchurch, New Zealand, Nov. 26-28, 2008.
[39] S. M. Lajevardi and M. Lech, "Averaged gabor filter features for facial expression recognition," in *Proc. DICTA*, Canberra, Australia, Dec. 1-3, 2008.

Abas Albahri received his BEng in Electrical and Electronic Engineering from the Sirt University, Libya in 2003. He was awarded his Master's degree in the Control System and Measurements from the Department of Electrical and Computer Engineering, Academy of Post Graduate Studies Tripoli, Libya in 2006. Currently, Abas is a PHD candidate at the School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia.

Margaret Lech received the M.S. degree in Physics from the Maria Curie-Skłodowska University, Poland, and the Ph.D. degree in Electrical Engineering from the University of Melbourne, Australia. She is currently a Senior Lecturer at the School of Electrical and Computer Engineering, RMIT University, Australia. Her research interests include psychoacoustic, speech and image processing, system modelling, and optimization.

Eva Cheng received her PhD degree in Telecommunications Engineering from the University of Wollongong, Australia. She is currently a Lecturer in the School of Electrical and Computer Engineering at RMIT University, Australia. Her research interests include areas in multimedia signal processing such as 3D video/audio recording and reproduction, computer vision, and speech/audio processing.