Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 231–240

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Investigating the effects of lossy compression on age, gender and alcoholic information in EEG signals

Binh Nguyen[a], Wanli ma[a], Dat Tran[a,*]

[a]*Faculty of Science and Technology, University of Canberra, ACT 2601, Australia*

Abstract

The age and gender information extracted from Electroencephalogram (EEG) has been used in various applications which are allocating a person to age and gender groups, identifying or authenticating a person and improving brain-computer interface systems. Besides this, the EEG-based automatic recognition of alcoholics greatly supports to the psychiatrists. However, one of the major challenges when using EEG is about storing and transmitting a huge amount of EEG data, leading to the need of using compression. Although lossy compression techniques give much higher compression ratio (CR) than lossless ones, they introduce the loss of information including the age, gender and alcoholic information in the reconstructed signals, which may reduce the performance of EEG-based age, gender and alcoholic recognition systems significantly. In this paper, the impact of lossy compression on the age, gender and alcoholic information extracted from EEG signals is examined in detail with different feature extraction and machine learning techniques. Our experimental results indicate that with an appropriate feature extraction technique, we could minimize the information loss in EEG compression and maintain the high performance of EEG-based age, gender and alcoholics recognition systems.

Keywords: EEG; EEG lossy compression; age and gender information; EEG-based age and gender recognition systems.

1. Introduction

Electroencephalogram (EEG) has been commonly used to extract the diagnostic information of seizure, dementia, and psychiatric disorders to analyse and detect various brain-related diseases [11]. Moreover, EEG has also been used as a new type of biometric for person recognition systems thanks to advantages such as universal and difficult to fake and attack, compared to conventional ones like fingerprint, face and voice [14, 22, 25]. Recently, extracting age, gender and alcoholic information on EEG signals has been investigated and EEG-based age, gender and alcoholics recognition systems have been proposed [12, 13, 21, 23, 24].

* Corresponding author. Tel.: +61-2-6201-2394.
E-mail address: dat.tran@canberra.edu.au

10.1016/j.procs.2019.09.178

In terms of EEG-based age and gender recognition, two common features, namely autoregressive (AR) and power spectral density (PSD), and paralinguistic features were used with the classifier of Support Vector Machine (SVM) to classify age and gender information in [21, 23]. In addition, entropy measurement methods and a range of classifiers like K-Nearest Neighbors (KNN), Logistic Regression (LR), SVM, Random Forest (RF), Quadratic Discriminant Analysis (QDA), and Decision Tree (DT) were proposed for gender recognition using EEG signals in [12]. Results from [12, 21, 23] stated that EEG signals contain rich age and gender information and the age and gender recognition systems based on EEG signals obtain high accuracies. The age and gender information extracted from EEG signals is very useful in various applications. For example, this information could be used to allocate a person to age or gender groups based on EEG characteristics of that person. Besides this, the age and gender information could be used to index EEG data for searching, identify or authenticate a person based on his/her EEG signals, and improve accuracy of brain-computer interface systems [21].

With respect to EEG-based alcoholics recognition, Palaniappan *et al*. proposed to use features extracted from Visual Evoked Potential (VEP) and classifiers of SVM and KNN for classifying the alcoholics [24]. In addition, two classifiers, in particular SVM and Multilayered Percepton (MLP), were used with features extracted from the power spectrum of the Haar mother wavelet to separate alcoholic signals from non-alcoholic ones in [13]. These proposed methods gave accurate recognition results over 94%, which considerably diminish the requirement of psychiatrists to classify the alcoholics.

Hence, EEG signals contain useful information on age, gender and alcoholics and using EEG signals in age, gender and alcoholics recognition is a promising research direction.

There are four factors that affect to the size of EEG data, which are the number of channels, the amount of time, bits resolution and sampling frequency. The number of channels can surpass 256 while the time can be several hours or even several months [15]. Although the sampling frequencies are normally between 256 Hz and 512 Hz, they may be up to 2000 Hz in some specific applications [27]. In addition, the more EEG signals are used, the better performance and reliability that EEG-based applications could obtain. As a result, a huge amount of EEG data need to be transmitted and stored, which is one of the major challenges in the usage of EEG signals. Therefore, using compression techniques is necessary.

There are two kinds of EEG compression techniques, namely lossless and lossy. Lossless compression ensures no loss of information in the reconstructed data, while lossy causes a loss in the recovered data. Thanks to obtain much higher compression ratio (CR) than lossless, lossy techniques are still widely studied and used. Currently, EEG lossy compression can be classified into four groups, namely Wavelet-based, Filter-based, Predictor, and Other [5]. Study in [5] also pointed out that most of lossy techniques are transformation-based, for example, Wavelet-based, and Discrete Wavelet Transform - Set Partitioning in Hierarchical Trees (DWT-SPIHT) and its extended versions are one of the best lossy compression. Moreover, Discrete Wavelet Transform - Adaptive Arithmetic Coder (DWT-AAC) is reported in [20] that it achieves good results compared to some recent compression techniques.

The use of EEG lossy compression to reduce the size of data is necessary, and hence the impact of lossy compression techniques on EEG-based applications should be investigated. With regard to EEG-based seizure recognition systems, Higgins *et al*. and Nguyen *et al*. stated that applying lossy compression is feasible and has the advantage compared to using lossless ones [10, 16, 18]. For instance, the seizure recognition accuracy remains at 90% when EEG data are compressed and then recovered by DWT-AAC at CR of 24 [18]. In consideration of EEG-based user authentication system, Nguyen *et al*. pointed out that the system accuracy is unchanged if the information loss in the recovered signals after using lossy compression is not greater than 11% [19]. As far as EEG-based person identification system is concerned, studies in [16, 17] examined the effects of lossy compression on this system. Results pointed out that although the system performance is reduced by lossy compression, the identification rate still keeps good enough (above 90%) when CR less than 25. In contrast, the impact of lossy compression on the age, gender and alcoholic information extracted from EEG signals has not been examined.

Two research questions here are 1) What is the impact of lossy compression on the age, gender and alcoholic information extracted from EEG signals? 2) Is it feasible for EEG-based age, gender and alcoholics recognition systems to use reconstructed signals processed by lossy compression? and if it is, what information loss can be tolerated? Our study aims to answer these questions.

The rest of the paper is structured as follows. Sections 2 and 3 outline the background information related to EEG lossy compression and EEG-based age, gender and alcoholics recognition systems, correspondingly. Section 4

presents experiment conditions, followed by the results and discussion in Section 5. The conclusion is presented in Section 6.

2. EEG lossy compression

This section provides a brief information of two lossy techniques, namely DWT, DWT-AAC, DWT-SPIHT, and performance measures.

2.1. Discrete Wavelet Transform (DWT)

Thanks to working on both time and frequency domains, DWT is suitable for non-stationary signals like EEG. A signal is decomposed by DWT into a set of basic functions, known as mother wavelets. Besides this, the high-pass and low-pass filters are used to filter each signal into approximation coefficients (low frequencies) and detail coefficients (high frequencies) respectively. The high-pass filter uses wavelet functions, while the low-pass ones uses the scaling functions. An alternative representation of the original signals is given by the approximation and detail coefficients. The reconstructed signals can be recovered by using Inverse DWT [1, 29].

2.2. DWT-AAC

DWT-AAC is an EEG lossy compression technique that proposed by Nguyen et al. [20]. In the compression process, EEG signals are decomposed by a DWT into sub-bands that contain DWT decomposition coefficients. Afterwards, those coefficients are quantised and then thresholded, which creates the binary significance map and indices of significant coefficients. Finally, both the binary significance map and indices of significant coefficients are coded by an AAC to generate the compressed signals. An inverse process is employed to decompress the signals.

Putting the thresholding component after quantisation helps DWT-AAC to control the impact of using static threshold values on the fidelity of signals. Moreover, using AAC instead of Arithmetic Coder or Huffman, and coding both the binary significance map and indices of significant coefficients improve the compression performance of DWT-AAC.

2.3. DWT-SPIHT

SPIHT is originally proposed for image compression by Said and Pearlman [26]. Its core principles are based on the embedded zerotree wavelet (EZW) introduced by Shapiro [28]. The wavelet coefficients produced by using DWT are put into the spatial orientation trees, the sorting and then refinement procedures where the relationship between wavelet coefficients in different scales is exploited efficiently to improve the performance of SPIHT.

2.4. Performance Measures

The widely used metrics to measure the performance of compression algorithms are compression ratio (CR), percentage root-mean-square difference (PRD).

CR is defined as the ratio between the number of bits of the original signals (L_{org}) and the compressed signals (L_{comp}):

$$CR = \frac{L_{org}}{L_{comp}} \tag{1}$$

PRD is used to evaluate the distortion between the original and recovered signals. It is defined as:

$$PRD = \sqrt{\frac{\sum_{i=1}^{N}(x_{org}[i] - x_{rec}[i])^2}{\sum_{i=1}^{N}(x_{org}[i])^2}} \times 100\% \tag{2}$$

where $x_{org}[i]$ and $x_{rec}[i]$ represent the original and reconstructed EEG signals correspondingly and N is the number of samples.

3. EEG-based age, gender and alcoholics recognition systems

The background information of EEG-based age, gender and alcoholics recognition systems will be presented in this section, which are system introduction, feature extraction and classifier SVM.

3.1. System Introduction

There are two phases in EEG-based age, gender and alcoholics recognition systems, namely training and testing. In training phase, EEG signals are captured and then pre-processed before extracting features. AR, PSD and paralinguistic features are used to build age, gender and alcoholics models using the classifier of SVM. In testing phase, the user is required to provide his/her EEG signals that he or she did in training phase. Similarly, the signals are pre-processed to extract features, and then these features are fed into the classifier to classify age, gender and alcoholics. In this research, AR, PSD and paralinguistic features, and SVM will be used for all age, gender and alcoholics recognition systems. This is because AR, PSD and SVM have been widely used in EEG-based pattern recongition systems [27] including age and gender recognitions [21, 23], whilst the paralinguistic feature gave high accuracy in age, gender and person recognition [21, 22, 23].

We investigate the impact of lossy compression on the age, gender and alcoholic information through studying the recognition performances using EEG data without and with compression at different levels of CR.

3.2. Feature Extraction

3.2.1. AR and PSD Features

The AR model of signals can be used for a single channel of EEG and each sample is defined to be linearly related to a number of its previous samples [27].

$$y(m) = -\sum_{j=1}^{p} a_j y(m-j) + x(m) \tag{3}$$

where $y(m)$ is EEG sample, a_j, $j = 1, 2, .., p$ are the linear parameters, m denotes the discrete sample time, and $x(m)$ is the noise input.

PSD is a function of frequency that shows where frequencies energy variations are strong or weak. PSD can be defined as the discrete time Fourier transform (DTFT) of the covariance sequence [31]:

$$S(w) = \sum_{l=-\infty}^{\infty} R(l) e^{-iwk} \tag{4}$$

where $R(l)$ is the autocorrelation function and defined as:

$$R(l) = E\{s(t)y^*(t-l)\} \tag{5}$$

where $s(t)$ is the discrete time signal $\{s(t); t = 0, \pm 1, \pm 2, \ldots\}$

3.2.2. Paralinguistic Features

EEG is a time-varying signal. Although EEG signals are non-stationary in a long period, they are considered as 'quasi-stationary' if the time window is sufficiently short. Based on this property, Nguyen et al. proposed to use the paralinguistic feature extraction method to extract brain wave features from EEG signals [22]. In this paper, an open-source openSMILE [7] was used to extract paralinguistic features including Mel-Frequency Cepstral Coefficients (MFCCs), Log filter-bank powers (LFBP), and Line spectral pairs (LSP). The range of frequencies was set between 1 to 300 Hz.

3.3. Support Vector Machine (SVM)

The training data is labeled as $\{x_i, y_i\}, i = 1, \ldots, n, y_i \in \{-1, 1\}, x_i \in R^d$, where n is the number of feature vectors and d is dimension. SVM uses C-Suport Vector Machine to construct an optimal hyperplane $f(x)$ [3], which can be used for classification, regression, or other tasks.

$$f(x) = w^T \Phi(x) + b \tag{6}$$

where w is the normal vector, Φ is a transformation from input space to feature space, and b is a constant. The optimisation problem is as follows

$$minimize \quad \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} \xi_i \tag{7}$$

$$subject\ to \quad y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n \tag{8}$$

where C is a parameter and $\xi_i, i = 1, \ldots, n$ are non-negative slack variables.

In the testing phase, SVM is used to compute the sign of

$$f(x) = \sum_{i}^{N_s} \alpha_i y_i \Phi(s_i)^T \Phi(x) + b = \sum_{i}^{N_s} \alpha_i y_i K(s_i, x) + b \tag{9}$$

where $s_i, i = 1, \ldots, n$ are support vectors.

4. Experiment Conditions

4.1. EEG datasets

Three public datasets were used for this research. Particularly, CHB MIT Scalp EEG Dataset [8] (DS1) contains 10 records of 10 people (4 males and 6 females) with the age from 2 to 19 and EEG signals were sampled at 256 Hz with 16-bit resolution. TUH EEG Corpus Dataset [9] (DS2) includes signals of 20 people (10 males and 10 females) with the age from 19 to 76. The sampling frequency and bits resolution are 256 Hz and 16, respectively. Finally, Alcoholism Large EEG Dataset (Test Data) [2] (DS3) contains EEG recordings of 10 alcoholics and 10 controls sampled at 256 Hz with 16-bit resolution.

4.2. Compression parameters

DWT was set up to run with 5 levels of biorthogonal 4.4 for both DWT-ACC and DWT-SPIHT because this mode has already achieved wide acceptance for use in compression algorithms [6]. Besides this, 6 bits uniform quantisation was used in DWT-AAC thanks to its high performance compared to some other techniques [20].

For both EEG lossy compression techniques, the signals were compressed and recovered at a range of different CRs, which helps to evaluate age, gender and alcoholics recognition performances with increasing compression.

4.3. Parameters of feature extraction and classifier

The recordings of each subject from DS1 and DS2 were split into 15s epochs for feature extraction. The 8 channels chosen were C3, C4, Cz, P3, P4, Pz, O1 and O2 for both datasets. The feature vectors were labelled for 3 age groups and 2 gender ones, resulting to 6 groups. Particularly, the children age range is 2-9, teenager age range is 10-18, and adult is 19 and over in DS1. Six groups are children male, children female, teenager male, teenager female, adult male and adult female. For DS2, 3 age groups are young (19-34), middle (35-54) and elderly (55-76), creating 6 classes which are young male, young female, middle male, middle female, elderly male and elderly female. Regarding to DS3, 19 channels were selected, namely FP1, FP2, F7, F8, FZ, F4, F3, T8, T7, CZ, C3, C4, P3, P4, PZ, P8, P7, O2

and O1. The signals of each trial (1s) of each subject were split for extracting features. Two labelled groups were alcoholics and non-alcoholics.

The spectral power in 2 Hz frequency bins from 1 to 30 Hz was computed for each channel and 21 AR coefficients of the 21th-order AR model were extracted using Burg's lattice-based method. For PSD features, the Welch's method using periodogram was used to create 12 frequency components in the band 8-30 Hz. Two-third of data were used for cross validation training whilst one-third for testing, which is applied to both datasets. Linear SVM classifier [4] was used to train models in 3-fold cross validation with parameter C ranging from 1 to 1000 in 5 steps. The accuracy rates of age and gender recognition were calculated from the confusion matrix of 6-classes recognition experiment in the test phase by summing the predictions over the desired groups. Features for training and testing were randomly selected ten times, and the final experimental result is the average of these ten results.

5. Results and Discussion

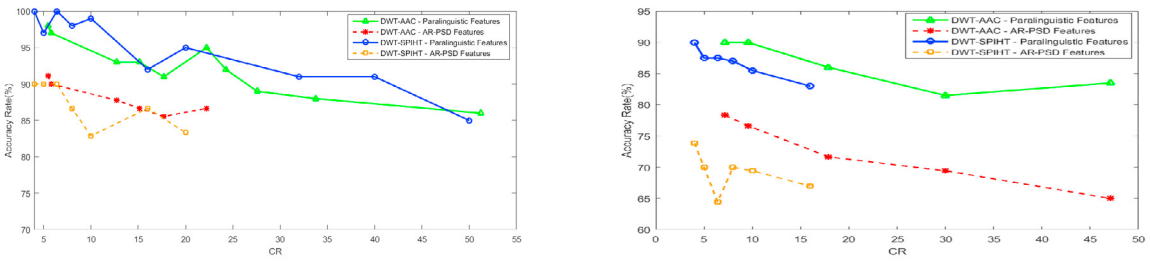
5.1. Age recognition performance with increasing lossy compression and maximising information loss

Table 1: Performance of age recognition using original EEG signals

Dataset	**Accuracy (%)** *Paralinguistics*	**Accuracy (%)** $AR - PSD$
DS1	97	97
DS2	93.8	81.6

Table 1 shows the results of the age recognition system using original EEG signals from DS1 and DS2, while Figures 1a and 1b plot the accuracy rates versus CRs using DWT-AAC and DWT-SPIHT. The performance of age recognition system reduces when compression is increased. On DS1, for instance, at CR of 6.4 processed by DWT-SPIHT, the accuracy rate is 100% when using paralinguistic features, compared to only 91% at CR of 32, as seen in Figure 1a. On DS2, for example, the system performance with AR - PSD features gives 78.3% when using DWT-AAC at CR of 7.2, while this figure is 69.5% when CR reaches 30, as shown in Figure 1b.

Additionally, it can be seen that using paralinguistic features obtains higher performances of age recognition than using AR-PSD ones. Particularly, at the same CR of 12.7 processed by DWT-AAC on DS1, the accuracy rate is 93% when using paralinguistic features while it is only 87.7% if using AR-PSD ones. Moreover, the age recognition rate is 90% when using paralinguistic features at CR of 4 processed by DWT-SPIHT on DS2, compared to only 73.8% if using AR-PSD at the same value of CR. Furthermore, using originals signals on DS2 with paralinguistic features, the system performance is 93.8%, whilst this figure is only 81.6% if using AR-PSD ones.

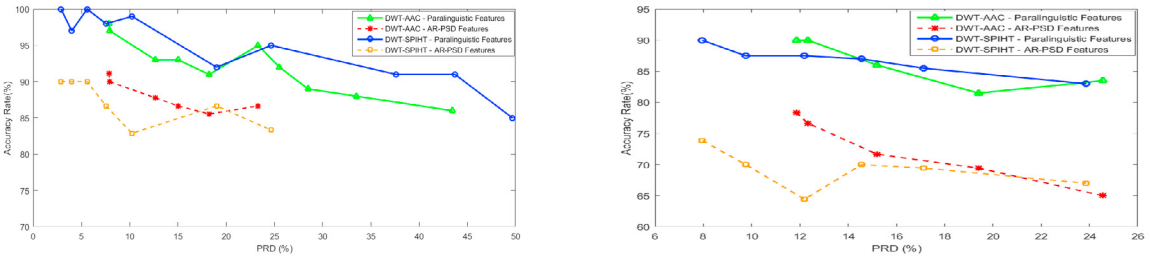


(a) DS1 (b) DS2

Fig. 1: Age recognition rate versus CR for DWT-AAC and DWT-SPIHT.

One of the purposes of this research is to determine how much information loss is tolerated if the EEG lossy compression does not affect significantly to the performance of EEG-based age recognition system. Studies in [10][17]

pointed out that a percentage greater than 90% is considered very good performing classifier. Hence, 90% is also considered very good performance for age recognition and 90% accuracy rate is used as a threshold limit for compression.



(a) DS1 (b) DS2

Fig. 2: Age recognition rate versus PRD for DWT-AAC and DWT-SPIHT.

As shown in Figure 2a showing the age recognition performances versus PRDs on DS1, and referring back to Figure 1a, for DWT-AAC, a 90% accuracy rate corresponds to CR of 26 and PRD of 27% when using paralinguistic features, while they are 6 and 8% if using AR - PSD. For DWT-SPIHT, CR and PRD are 41 and 44% correspondingly when using paralinguistic features, compared to 6.4 and 6% respectively if using AR-PSD ones.

As seen in Figure 2b showing the age recognition performances versus PRDs on DS2, and referring back to Figure 1b, there is no accuracy rate that is equal or greater than 90% if using AR-PSD features for both lossy techniques. Conversely, when using paralinguistic featues, at 90% accuracy rate, CR and PRD of DWT-AAC are 9.5 and 13% respectively, while those figures are 4 and 8% for DWT-SPIHT correspondingly.

5.2. Gender recognition performance with increasing lossy compression and maximising information loss

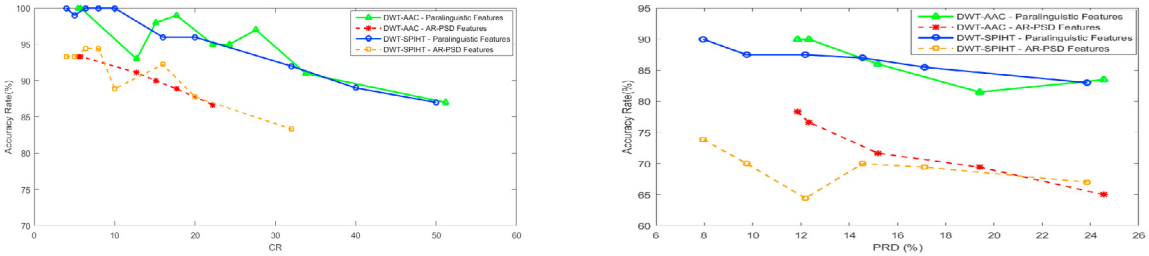
Table 2: Performance of gender recognition using original EEG signals

Dataset	**Accuracy (%)** *Paralinguistics*	**Accuracy (%)** $AR-PSD$
DS1	100	97.7
DS2	94.5	87.7

Table 2 shows the performances of gender recognition using original EEG data on DS1 and DS2. In addition, Figures 3a and 3b illustrate the plots of gender accuracy rates versus corresponding CRs for both DWT-AAC and DWT-SPIHT on DS1 and DS2, respectively.

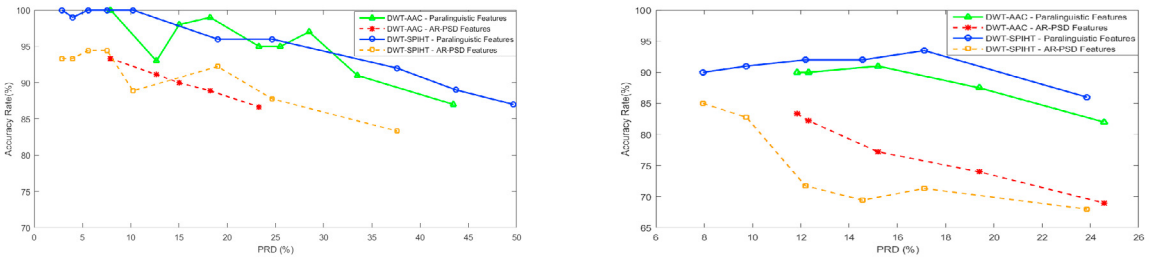
Similar to age recognition, the performance of gender recognition reduces gradually when CR increases. On DS1, for example, at CRs around 6 of both lossy techniques with paralinguistic features, accuracy rates are 100%, compared to 95% when CRs around 25. Similarly, the accuracy rates using AR - PSD features are 93% when CRs of both lossy techniques are 6, whilst the figures are around 87% when CRs of 20. On DS2, for instance, using DWT-AAC at CR of 7.5, the accuracy rates with paralinguistic and AR-PSD features are 90% and 83% correspondingly, while these figures are 87.5% and 75% when CR of 30, respectively.

Besides this, the gender recognition using paralinguistic features obtains higher performances than that using AR - PSD ones for both original and recovered EEG signals at different CRs. In particular on DS1, the gender recognition performance is 100% when using original signals with paralinguistic features, compared to 97.7% for AR - PSD. Moreover, using DWT-AAC at CR of 6, the paralinguistic features give the accuracy rate of 100%, compared to 93% when using AR - PSD ones, as seen on Figure 3a. Another example on DS2 is that the performance using original signals is 94.5% when using paralinguistic features whilst it is 87.7% if using AR - PSD ones, as shown in Table 2. Furthermore, the accuracy rate using paralinguistic features is 92% at CR of 8 processed by DWT-SPIHT, while it is only 69.4% when using AR - PSD at the same CR.



(a) DS1 (b) DS2

Fig. 3: Gender recognition rate versus CR for DWT-AAC and DWT-SPIHT.



(a) DS1 (b) DS2

Fig. 4: Gender recognition rate versus PRD for DWT-AAC and DWT-SPIHT.

In similar, the accuracy rate of 90% is selected as a threshold limit for compression in the gender recognition system. Figure 4a shows the plots of gender accuracy rates versus corresponding PRDs for DWT-AAC and DWT-SPIHT on DS1. Referring back to Figure 3a, at an accuracy rate of 90% when using paralinguistic features, CR and PRD of DWT-AAC are 39 and 37% respectively, while those figures for DWT-SPIHT are 39 and 43% correspondingly. For AR - PSD features at the accuracy rate of 90%, CR and PRD of DWT-AAC are 15 and 15% respectively, whilst those of DWT-SPIHT are 17 and 22% correspondingly.

Figure 4b demonstrates the plots of gender accuracy rates versus corresponding PRDs for DWT-AAC and DWT-SPIHT on DS2. Referring back to Figure 3b, it can be seen that only paralinguistic features at some different CRs give the accuracy rate equal or greater than 90%. In contrast, no values of CR for both lossy techniques is recorded in which the gender recognition performance with AR-PSD features is equal or greater than the threshold of 90%. For paralinguistic features, for instance, at 90% accuracy rate, DWT-AAC has CR of 21 with PRD of 16%, compared to 12 and 20% for DWT-SPIHT correspondingly.

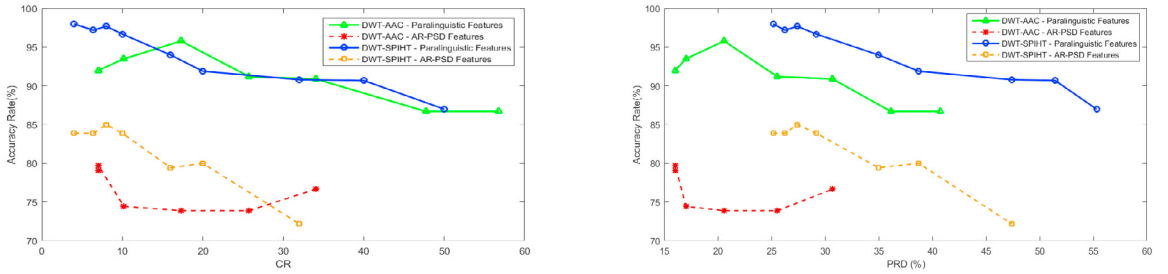
5.3. Alcoholics recognition performance with increasing lossy compression and maximising information loss

Table 3: Performance of gender recognition using original EEG signals

Dataset	**Accuracy (%)** *Paralinguistics*	**Accuracy (%)** $AR-PSD$
DS3	90.5	87.22

Table 3 illustrates the results of the alcoholics recognition system using original EEG signals on DS3, whilst Figure 5a plots the accuracy rates versus CRs using DWT-AAC and DWT-SPIHT. Generally, the performance of alcoholics recognition system diminishes when augmenting compression. For example, at CR of 6.4 processed by DWT-SPIHT,

the accuracy rate using paralinguistic features is 97%, compared to only 91% at CR of 31. Similarly, the system performance with AR - PSD features gives 80% when using DWT-AAC at CR of 7, while this figure is 73.5% when CR reaches 26. Furthermore, the alcoholics recognition using paralinguistic features gives much higher performances than that using AR-PSD ones for both original signals and signals processed by two lossy compression techniques. For original signals, the accuracy rate is 90.5% when using paralinguistic features, compared to 87.2% if using AR-PSD ones, as seen in Table 3. At the same CR of 6.4 processed by DWT-SPIHT, for example, using paralinguistic features obtains 97% accuracy rate, while it is only 84% when using AR-PSD ones, as shown in Figure 5a.



(a) Acurracy rate versus CR.

(b) Acurracy rate versus PRD.

Fig. 5: Alcoholics recognition using reconstructed signals processed by DWT-AAC and DWT-SPIHT.

Similar to age and gender recognition systems, 90% of the accuracy rate is used to limit compression in the alcoholics recognition system. From Figure 5b showing the accuracy rates versus PRDs and referring back to Figure 5a, only using paralinguistic features that gives the accuracy rates equal or greater than 90% for both DWT-AAC and DWT-SPIHT. In particular, CR and PRD of DWT-SPIHT reach 41 and 52% respectively at 90% accuracy rate, while those figures are 36 and 32% correspondingly for DWT-AAC.

To sum up, from the above results, it can say that lossy compression do have an impact on the age, gender and alcoholic information extracted from reconstructed EEG signals as it causes an information loss. This is shown via the performance of EEG-based age, gender and alcoholics recognition systems as it decreases gradually when increasing compression. This is because the more compression, the more information loss in the recovered signals including the age, gender and alcoholic information, which results in the performance degradation of the age, gender and alcoholics recognition systems. Furthermore, the recognition system using paralinguistic features gives a higher performance than that using AR - PSD ones. As presented in Sections 3.2 and 4.3, the frequency ranges for paralinguistic and AR-PSD features are 1-300 Hz and 1-30 Hz respectively. Basically, when compression, DWT-SPIHT and DWT-AAC will keep significant coefficients at low frequencies (with higher energy) and remove insignificant coefficients at high frequencies (with lower energy) [20, 26]. However, EEG signals are complex, non-stationary and the signal energy spreads over different frequencies [27]. As a result, significant coefficients at high frequencies, which may contain the age, gender and alcoholic information, still exist after compression. AR-PSD features that focus on low frequencies may miss valuable information at high frequencies, which degrades the performance of recognition systems, compared to those using paralinguistic features. Besides this, Srinivasan *et al*. stated that the CR of lossless compression of EEG signals only achieves between 2 and 3 [30]. Hence, in case of using paralinguistic features, using lossy compression take the advantages, compared to using lossless ones.

6. Conclusion

This paper has investigated the impact of lossy compression on the age, gender and alcoholic information extracted from EEG signals and our experimental results have answered the above-mentioned two research questions as follows: 1) Lossy compression techniques do have the impact on age, gender and alcoholic information by reducing age, gender and alcoholics recognition performances with increasing compression. Particularly, the recognition system will be affected more significantly if it uses AR - PSD features, compared to using paralinguistic ones. 2) In case of using paralinguistic features, it is feasible to apply lossy technique to EEG-based age, gender and alcoholics recognition

systems. When 90% is considered as a threshold for classifier, in the best case, the compression can be tolerated at CR up to 41 (PRD of 44%) for age recognition, at CR up to 39 (PRD of 43%) for gender recognition, and at CR up to 41 (PRD of 52%) for alcoholics recognition.

References

[1] Barford, L.A., Fazzio, R.S., Smith, D.R., 1992. An introduction to wavelets. Citeseer.

[2] Begleiter, H., 1999. Eeg database. URL: http://kdd.ics.uci.edu/databases/eeg/eeg.data.html.

[3] Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, 121–167.

[4] Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2, 27.

[5] Dao, P.T., Li, X.J., Do, H.N., 2015. Lossy compression techniques for eeg signals, in: Advanced Technologies for Communications (ATC), 2015 International Conference on, IEEE. pp. 154–159.

[6] Daou, H., Labeau, F., 2014. Dynamic dictionary for combined eeg compression and seizure detection. IEEE journal of biomedical and health informatics 18, 247–256.

[7] Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, ACM. pp. 1459–1462.

[8] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E., 2000. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 101, e215–e220.

[9] Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., Tobochnik, S., 2014. The tuh eeg corpus: A big data resource for automated eeg interpretation, in: Signal Processing in Medicine and Biology Symposium (SPMB), 2014 IEEE, IEEE. pp. 1–5.

[10] Higgins, G., McGinley, B., Faul, S., McEvoy, R.P., Glavin, M., Marnane, W.P., Jones, E., 2013. The effects of lossy compression on diagnostically relevant seizure information in eeg signals. IEEE journal of biomedical and health informatics 17, 121–127.

[11] Hill, D., 1958. Value of the eeg in diagnosis of epilepsy. British medical journal 1, 663.

[12] Hu, J., 2018. An approach to eeg-based gender recognition using entropy measurement methods. Knowledge-Based Systems 140, 134–141.

[13] Kousarrizi, M.R.N., Ghanbari, A.A., Gharaviri, A., Teshnehlab, M., Aliyari, M., 2009. Classification of alcoholics and non-alcoholics via eeg using svm and neural networks, in: Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on, IEEE. pp. 1–4.

[14] Marcel, S., Millán, J.d.R., 2007. Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. IEEE transactions on pattern analysis and machine intelligence 29, 743–752.

[15] Marsan, C.A., Zivin, L., 1970. Factors related to the occurrence of typical paroxysmal abnormalities in the eeg records of epileptic patients. Epilepsia 11, 361–381.

[16] Nguyen, B., Ma, W., Tran, D., 2018a. Impact of lossy data compression techniques on eeg-based pattern recognition systems, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. pp. 2308–2313.

[17] Nguyen, B., Ma, W., Tran, D., 2018b. A study of combined lossy compression and person identification on eeg signals, in: The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications, Springer. pp. 449–458.

[18] Nguyen, B., Ma, W., Tran, D., 2018c. A study of combined lossy compression and seizure detection on epileptic eeg signals. Procedia Computer Science 126, 156–165.

[19] Nguyen, B., Nguyen, D., Ma, W., Tran, D., 2017a. Investigating the possibility of applying eeg lossy compression to eeg-based user authentication, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE. pp. 79–85.

[20] Nguyen, B., Nguyen, D., Ma, W., Tran, D., 2017b. Wavelet transform and adaptive arithmetic coding techniques for eeg lossy compression, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE. pp. 3153–3160.

[21] Nguyen, P., Tran, D., Huang, X., Ma, W., 2013a. Age and gender classification using eeg paralinguistic features, in: Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on, IEEE. pp. 1295–1298.

[22] Nguyen, P., Tran, D., Huang, X., Sharma, D., 2012. A proposed feature extraction method for eeg-based person identification, in: International Conference on Artificial Intelligence.

[23] Nguyen, P., Tran, D., Vo, T., Huang, X., Ma, W., Phung, D., 2013b. Eeg-based age and gender recognition using tensor decomposition and speech features, in: International conference on neural information processing, Springer. pp. 632–639.

[24] Palaniappan, R., 2006. Improved automated classification of alcoholics and non-alcoholics. Information Technology 2, 182–186.

[25] Riera, A., Soria-Frisch, A., Caparrini, M., Grau, C., Ruffini, G., 2008. Unobtrusive biometric system based on electroencephalogram analysis. EURASIP Journal on Advances in Signal Processing 2008, 18.

[26] Said, A., Pearlman, W.A., 1996. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Transactions on circuits and systems for video technology 6, 243–250.

[27] Sanei, S., Chambers, J.A., 2007. EEG signal processing. Wiley-Interscience.

[28] Shapiro, J.M., 1993. Embedded image coding using zerotrees of wavelet coefficients. IEEE Transactions on signal processing 41, 3445–3462.

[29] Skodras, A.N., 2003. Discrete Wavelet Transform: An Introduction.

[30] Srinivasan, K., Reddy, M.R., 2010. Efficient preprocessing technique for real-time lossless eeg compression. Electronics Letters 46, 26–27.

[31] Stoica, P., Moses, R.L., et al., 2005. Spectral analysis of signals .