Pertanika J. Sci. & Technol. 27 (2): 737 - 752 (2019)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Real and Complex Wavelet Transform Approaches for Malaysian Speaker and Accent Recognition

Rokiah Abdullah[1]*, Hariharan Muthusamy[2], Vikneswaran Vijean[1], Zulkapli Abdullah[3] and Farah Nazlia Che Kassim[1]

[1]School of Mechatronic Engineering, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 UniMAP, Arau, Perlis, Malaysia
[2]Department of Biomedical-Engineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur 603 203 Kancheepuram District, Tamil Nadu, India
[3]Pusat Kejuruteraan and Inovasi, Universiti Malaysia Perlis, 02600 UniMAP, Arau, Perlis, Malaysia

ABSTRACT

A new approach for speaker and accent recognition based on wavelets, namely Discrete Wavelet Packet (DWPT), Dual Tree Complex Wavelet Packet Transform (DT- CWPT) and Wavelet Packet Transform (WPT) based non-linear features are investigated. The results are compared with conventional MFCC and LPC features. k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM) classifier are used to quantify the speaker and accent recognition rate. The database for the research was developed using English digits (0~9) and Malay words. The highest accuracy for speaker recognition obtained is 93.54% while for accent recognition; it is 95.86% using Malay words. Combination of features for speaker recognition is obtained from ELM classifier is 98.68 % and for accent recognition is 98.75 % using Malay words.

ARTICLE INFO

Article history:
Received: 15 June 2018
Accepted: 2 November 2018
Published: 25 April 2019

Keywords: Accent recognition, Discrete Wavelet Packet (DWPT), Dual Tree- Complex Wavelet Packet Transform (DT-CWPT), Speaker recognition, Wavelet Packet Transform (WPT)

E-mail addresses:
rokiah@unimap.edu.my (Rokiah Abdullah)
hariharan.m@ktr.srmuniv.ac.in (Hariharan Muthusamy)
vikneswaran@unimap.edu.my (Vikneswaran Vijean)
zulkapli@unimap.edu.my (Zulkapli Abdullah)
nazlia@unimap.edu.my (Farah Nazlia Che Kassim)
* Corresponding author

INTRODUCTION

Biometric refers to the identification of humans by their characteristic or traits. It can be categorized into physiological versus

ISSN: 0128-7680
e-ISSN: 2231-8526

behavioural characteristic. Common physical characteristic are fingerprints, face and palm print. Typing rhythm, gait and voice are examples of behavioural characteristics which are often related to behaviour of a person. Some of the characteristics are unique to every individual as it varies from one person to another, even if they are twins (Jain & Sharma, 2013; Kinnunen & Li, 2010; Anand et al., 2012).

Many studies have investigated speech/speaker and accent recognition using voice signals. There are various well-known feature extraction techniques for extracting useful features from voice such as MFCC, LPC, Linear Predictive Cepstral Coefficients (LPCC). MFCC and LPC features are widely preferred for studies on speech, speaker and accents recognition.

MFCC and LPC methods transform the speech signal from time-based to frequency-based domain. In transforming the signals, time information is lost. Wavelet based analysis does not use a time-frequency region, but rather a time-scale region. The wavelets work by scaling properties. They are localized in time and frequency, permitting a closer connection between function being represented and their coefficients (Lee & Yamamoto, 1994).

Wavelets approaches have proven to be one of the promising techniques in applications such as infant cry classification, speech signals processing for pathological detection and voice access system to name a few (Oung et al., 2018; Lim et al., 2016; Johari et al., 2011).

Although there are many studies on Malay speech/speaker and accent recognition using wavelets, the difference in characteristics between different types of wavelet transforms are less explored (Yusnita et al., 2012; Almaadeed et al., 2015; Yadav & Bhalke, 2015; Pandiaraj & Kumar, 2015; Lei & Kun, 2017). Therefore, this study undertakes the task of evaluating the performance of non-linear features derived from the various wavelet based approach (DWPT, DT-CWPT and WPT) in predicting speaker and accent recognition. Conventional LPC and MFCC parameters are also derived, and combination of these features with non-linear entropies are also evaluated in an effort to identify new parameters that can contribute to overall best prediction rate for speaker and accent recognition. A new speech database consists of Malay words uttered by Malaysian speakers from three major races, namely Malay, Chinese and Indian were used. Since we are using the Malay words, it will not only give the advantages for Malaysian who speak the language but also for people who speak this language in the South- east Asia such as Indonesia, south Thailand and south Philiphines (Hanifa et al., 2017).

RELATED WORKS

This section describes a previous research works in speech/speaker and accent recognition area using wavelets. Yusnita et al. (2012) studied hybrid Discrete Wavelet Transform (DWT) feature space using uniform and dyadic extraction of Linear Predictive Coding (LPC) for accent classification using k-Nearest Neighbors (k-NN). The best classification

rate was 93.25 % with 32- and 21- dimension space for uniform and dyadic manner. The dyadic type DWT-LPC yielded an increase in classification rate by 9.28 % with respect to conventional LPC method. Almaadeed et al. (2015) proposed speaker identification using multimodal neural networks and wavelet analysis. This approach used multiple Neural Network (Probabilistic Neural Network (PNN), Radial Basis Function NN (RBF-NN) and General Regressive NN (GRNN) using wavelet-based selection method. The proposed system obtained 97.5% accurate with a 50 ms identification time. Performance tests conducted using the GRID database corpora had shown that this approach had faster identification time and greater accuracy, compared to traditional approaches, and it was applicable to real-time, text-independent speaker identification systems.

According to Yadav and Bhalke (2015), speaker identification system based on the wavelet transform which is DWT based MFCC and Traditional MFCC are used as a feature for speaker identification system. MFCC based DWT results show 85% accuracy & Traditional MFCC results show 80% accuracy.

Pandiaraj and Kumar (2015) discovered speaker identification system using DWT and Gaussian Mixture Model (GMM) used for classification. Daubechies wavelets were used and analysed using 8 levels of decomposition. The maximum accuracy of 83.3 % was achieved for the proposed method.

Lei and Kun (2017) researched on speaker recognition using Wavelet Packet Entropy (WPE), I- Vector and Cosine Distance Scoring (CDS) in noisy environment. Experimental results showed that WPE-I-CDS was robust in noisy environment compared with MFCC-I-CDS and Fused MFCC (FMFCC)-I-CDS. Based on the 94.36% of accuracy obtained, it was concluded that WPE using i-vector and CDS classifier was best suited for speaker recognition.

Rathor and Jadon (2017) proposed text independent speaker recognition using Wavelet Cepstral Coefficient (WCC) and Butter worth filter. The Wavelet Transform was used to find the frequency spectrum while WCC was used to capture the characteristic of the signal. The proposed method obtained 98.5% accuracy by using Butter worth filter. The authors concluded that the proposed method achieved a good performance in noisy environment.

Chelali and Djeradil (2017) had developed text dependent speaker recognition applied for Algerian Berber language using MFCC, delta MFCC, delta-delta MFCC, LPC and DWT. Identification rate for MFCC varied from 83% for the word "Tazalit" to 100 % for the word "Attas". LPC technique combined with DWT improved the efficiency of the system. The speaker recognition system improved the identification by 10 % compared with the classical MFCC and reduced identification time since the length was less than MFCC.

Motivated by previous studies, this study was undertaken to improve recognition rate of speaker and accent recognition using MFCC, LPC, DWPT, DT-CWPT and WPT based combined features. New database that contained English digits (0-9) and Malay words from

the major races in Malaysia; Malay, Chinese and Indian was constructed. The aims of this study are: (i) to compare performances of feature extraction methods for English digits and Malay words using speaker and accent database; and (ii) to study the performance of combined feature extraction methods using different classifiers.

METHODOLOGY

Database

The speech corpus was created from 39 male and female undergraduate students of University Malaysia Perlis from different races. Each speaker pronounced the English digits (0~9) and Malay words for 15 sessions. Every session consisted of predefined digit and Malay word organized randomly. The Malay words selected represented the six vowels of a, e, i, o, u, e' and had a combination of consonants and vowels in monosyllable and bi-syllable structure. The total speech samples were 12285 files. Table 1 and 2 summarize the database.

Experiment Setup

The experiment was conducted based on the methodology as presented in Figure 1. Speech signals were recorded with a sampling rate of 44 kHz and down-sampled to 16 kHz. Based on the Shannon sampling theorem, the 16 kHz sampling was enough to reconstruct an 8 kHz bandwidth signal (telephony speech bandwidth) maximum frequency (Gruhn et al., 2011). In Pre-processing stage, the speech signal was normalized and filtered, so that, only useful speech information was retained. In feature extraction process, MFCC, LPC, DWPT, DT-CWPT and WPT based features were extracted from the sampled speech signals. At this stage, all of the information necessary to distinguish speaker and accent was preserved. Configurable feature combination block selected which of the features to be used for accuracy calculation. It supported single feature or combined features extraction output. The accuracy was investigated for individual and combined features using k-NN, SVM and ELM classifier for Malaysian speaker and accent recognition. This process involved classifying the speech signal to determine whether the input speech matched any of learnt speech.

Feature Extraction

Mel Frequency Cepstral Coefficients (MFCC). The MFCC was introduced by Davis and Mermelstein (Saksamudre & Deshmukh, 2015) which was based on human hearing perceptions and cannot perceive frequencies over 1Khz. It is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. This can be represented mathematically as:

$$M(f) = 1125 * \log_e(1 + \frac{f}{700}) \quad (1)$$

In this work, MFCC is used to extract feature from input signal.

Frame size for the analysis was set to 512 sample points, such that the time period of the signal is 512/16000= 32ms. This is because a short period of time (20 – 40msec) speech signals are known to exhibit quasi-stationary behavior (Jain & Sharma, 2013). Hamming window was used to smooth the signal and make it more amendable for spectral analysis. Fast Fourier Transform (FFT) was applied to convert each frame of 256 samples from time domain into frequency domain. The Mel scale is based on pitch perception and triangular-shaped filter was used. Discrete Cosine Transform (DCT) is used to convert the log mel spectrum into time domain. The result of conversion is called MFCC and the set of coefficients is acoustic vector. Thirteen acoustic vectors were used to represent and recognize the voice characteristic of the speaker for this study.

Table 1

Malay word syllable structure

Word	Phoneme Sequence	Syllables	No. of syllable
Jam	/Jam/	CVC	1
Pas	/Pas/	CVC	1
Cap	/Cap/	CVC	1
Tol	/Tol/	CVC	1
Sen	/Sen/	CVC	1
Aku	/A-ku/	V-CV	2
Basi	/Ba-si/	CV-CV	2
Pulau	/Pu-lau/	CV-CVV	2
Rabu	/Ra-bu/	CV-CV	2
Jalan	/Ja-lan/	CV-CVC	2
Muka	/Mu-ka/	CV-CV	2

Table 2

Database details

Item	Description
Speakers	39
Session /speaker	15 times
Wordlist	1.Digit English (0~9) 2.Malay word (Jam, Pas, Cap, Tol, Sen, Aku, Basi, Pulau, Rabu, Jalan, Muka)

Table 2 *(Continue)*

Item	Description
Age	19- 24 years old
Race	Malay, Chinese, Indian
Microphone	Stereo microphone
Types of room	Controlled environment
Sampling frequency	16kHz
Audio file format	Wav

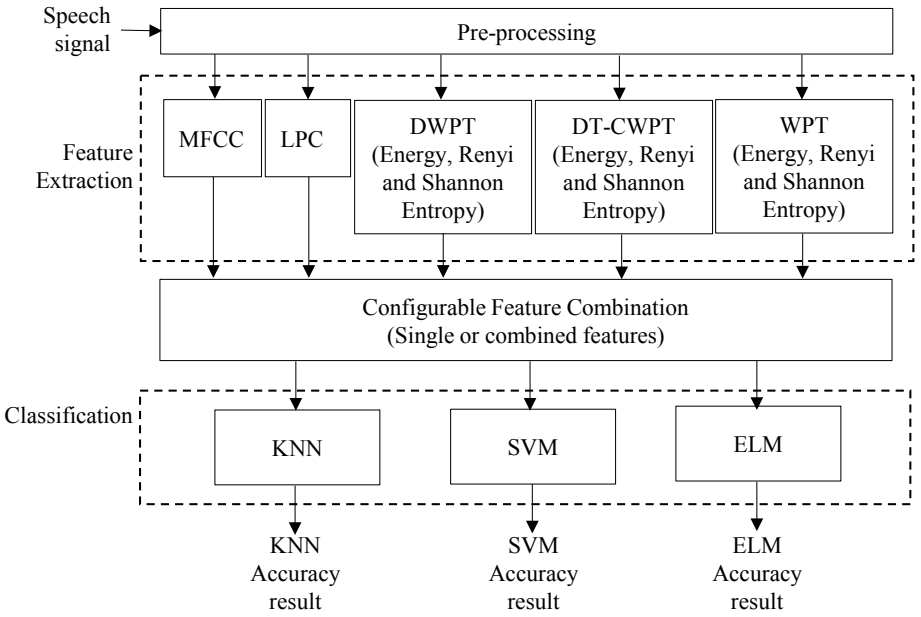

Figure 1. Flow chart of proposed methodology

Linear Predictive Coding (LPC). LPC analysis models the speech signal as a p- order autoregressive (AR) system which is a special case of all-pole IIR filter. The current value of the real-valued time series, is predicted based on past samples by minimizing the prediction error in the least squares sense (Paulraj et al., 2008). All the speaker data can be approximated to be a linear combination of past samples given by:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \qquad (2)$$

where ŝ(n) is the estimated sample, p is the order of the model, a_k is the linear predictive coefficients and s (n-k) is the previous speech sample. P orders range between 8-20 gives

good performances for recognition system. Prasanna et al., (2006) andYusnita et al., (2011) found that 16 orders gave a good result. Thus, in this study we set the orders of p as 16.

Discrete Wavelet Packet Transform (DWPT). DWPT is an extension of DWT, whereby all nodes in the tree structure are allowed to split further at each level of decomposition. In the DWT, each level is calculated by passing only the previous wavelet approximation coefficients through discrete-time low and high pass quadrature mirror filters. However, in the DWPT; both the detail and approximation coefficients are decomposed to create the full binary tree. Therefore, features can be generated based on approximation and detail coefficients at different levels to obtain more information (Zhang et al., 2015).

Dual Tree Complex Wavelet Packet Transform (DT-CWPT). DT-CWPT's shift invariance and directional selectivity provides an accurate measure of spectral energy at a particular location in space, scale and orientation (Lim et al., 2016). The DT-CWPT consisted of two DWPT operating in parallel on an input signal. The second wavelet packet filter bank was obtained by replacing the first stage filter $h_i^{(1)}(n)$ by $h_i^{(1)}$ (n -1) and by replacing by h_i'(n) for $i \in \{0,1\}$.

For the research, input speech signals were decomposed into 5 levels using DT_CWPT. Non-linear entropy features were extracted from each sub-band for the analysis, which produced 124 feature vectors. -

Wavelet Packet Transform (WPT). Wavelet Packet Transform (WPT) is an extension of Discrete Wavelet Transform and can be obtained by a generalization of the fast-pyramidal algorithm. For DWT decomposition procedure, signal is decomposed into lower frequency band (approximation coefficients) and higher frequency band (detail coefficients). For further decomposition, low frequency band is used and hence, DWT gives a left recursive binary tree structure. However, in Wavelet Packet Transform, lower and higher frequency bands are decomposed into two sub-bands. Therefore, wavelet packet gives a balanced binary tree structure (Johari et al., 2011). Forth order Daubechies wavelets were used for the analysis based on observations from works by (Lei & Kun, 2017; Bong et al., 2017) that demonstrated that this particular wavelet family was best suited for analysis of speech signals. Daubechies wavelet are found to be time invariant, computationally fast and has sharp filter transition bands (Cohen et al., 2006).

Classifier

k-Nearest Neighbor (k-NN). k-NN classifier is used to classify the English digits and Malay words. Due to its simple implementation and flexibility to feature/distance choices, k-NN is considered in this works. The k-NN classification system is a simple, supervised algorithm that employs lazy learning (Hariharan et al., 2012). The test samples are classified

based on majority of k-Nearest Neighbor's category. An object being classified to the class most common amongst its k nearest neighbors where k is a positive integer. The Euclidean Distance measure is used to calculate the closest members of the training set to test class being examined.

From this k-NN category, class label is determined by applying majority voting. Euclidean Distance is shown in Equation 3.

$$d_E[x,y] = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (3)$$

Normally, larger values of *k* can cause boundaries classes to be less distinct and will reduce the effect of noise on the classification (Liu et al., 2010). However, a lot of neighbors means neighbors that are far apart are also counted, which are irrelevant. Therefore, in this study, *k* values were varied between 1 and 10.

Support Vector Machine (SVM). For data classification, SVM is a supervised algorithm that can be used for two classes and multiclass recognition. It is based on principle of Structural Risk Minimization (SRM). It searches the best compromise between complexity of model and learning ability on the basis of limited sample information to obtain the best generalization ability. SVM generates an excellent performance which comes out from the fact that, SVM apply a linear algorithm to the data in a high dimensional space (Amami et al., 2015).

The parameters of the best C (Cost) and gamma (G) were optimized using Lib SVM Tool (Chang & Lin, 2011). SVM was chosen since it has a better generalization (less overfitting) and robust to noise.

Extreme Learning Machine (ELM). A new learning of single hidden layer feedforward networks (SLFNs), proposed by Huang et al (Cao et al., 2015). It has been used in various applications to overcome the slow training speed and overfitting problems of the conventional neural network learning algorithms (Sangeetha & Radha, 2013). The idea in ELM is that the weight of the hidden nodes and output nodes are randomly selected and analytically determined. ELM was chosen for having a better performance in learning efficiency and universal approximation capability. Moreover, it is a fast learning speed and good generalization performance. In this study the best value of the regularization coefficients of ELM classifier was found between -10 and 10.

RESULTS AND DISCUSSIONS

Table 3-6 show the results of MFCC, LPC, DT-CWPT, DWPT and WPT features for Malaysian speaker and accent recognition. The maximum recognition accuracies are highlighted in Table 3. From the highlighted results in Table 3, the highest accuracies for speaker recognition using English digits was 92.16 % and for Malay words was 93.54 % achieved using ELM classifier. It was found that the highest accuracies from speaker recognition was obtained from DT-CWPT features. This indicates that the percentage of recognition accuracies was improved using Wavelets from DT-CWPT features. Table 4 shows accuracy of accent recognition using English digits and Malay words. SVM classifier and MFCC achieved maximum accuracy rate of 94.48 % for accent recognition using English digits and for Malay words was 95.86%. The percentage shows that the recognition accuracies using MFCC, LPC, DT-CWPT, DWPT and WPT features are comparable.

Table 5 shows accuracy of speaker recognition using English digits and Malay words by combining the features. ELM has yielded higher recognition rate of about 98.09 % for English digits and 98.68 % for Malay words. Table 6 contains the results of accent recognition using English Digits and Malay words. It is observed that the maximum accuracy achieved from ELM classifier was 98.15% for English digits and for Malay words was 98.75 %.

From table 3-4, it can be summarized that in the accuracy results for MFCC, LPC and Wavelet based approach (DWPT, DT- CWPT and WPT), SVM performed better than ELM. Meanwhile for the combined features in Table 5 and 6, ELM gave better result. This is because combined features generate bigger data sets. ELM is well suited for solving big data and their solution is so rapidly obtained (Akusok et al., 2015). SVM with a greater number of samples will start to drop in terms of performance (Li & Yu, 2014). SVM has high algorithmic complexity and extensive memory requirements due to the use of a quadratic programming (Valyon & Horváth, 2003). From Table 3-4, in every experiment for speaker and accent recognition, ELM and SVM classifier showed a better performance than k-NN which was run separately. However, for combined features as can be seen in Table 5-6, ELM gave a slightly better result than SVM. From the results displayed in Tables 3-6, higher recognition rate was obtained using Malay words compared to English digits. It is because of vowels in the words have significantly more energy than consonants (Mohd Yusof et al., 2008). From a previous work in speaker/speech and accent identification/recognition, the accent identification researched by Yusnita et al., (2012) using hybrid DWT-LPC features and k-NN showed promising accuracy. The classification rate was 93.25 % compared than the conventional LPC while retaining the feature size. Adam et al (2013) reported an improved feature extraction method using Wavelet Cepstral Coefficients (WCC) recognized 26 English alphabets. The authors had found that WCCS showed comparable result with MFCCs. The best recognition was found from WCCs at level 5 of the DWT decomposition with a small difference of 1.19 % and 3.21 % when compared to MFCCs. Meanwhile Islam et al., (2016) proposed a new speaker identification system using 2-D neurograms constructed

from the responses of a physiologically-based computational model of the auditory. The identification score was found to be 93.5 % (40 dB), 93.5 % (60 dB) and 96.5 % (90 dB). Soon et al (2017) researched on speech recognition system for spoken English and Malay words from a group of Malay native speakers using DWT. Surface electromyogram (sEMG) employed to capture the speech and feature extraction was done in both temporal and time-frequency domains. The classification result showed that the Malay words ('satu', 'dua', 'tiga', 'empat', 'lima') gave a promising accuracy than English words ('one', 'two', 'three', 'four', 'five').

In this work, the accuracy for individual features was able to achieve 93.54 % (speaker) and 95.86 % (accent) while for the combined features, the result obtained were 98.68 % (speaker) and 98.75 % (accent). It is observed that the results are slightly better with previous works. This is because a variety of information inside features in DWPT, DT-CWPT and WPT contributes the promising accuracy. The results prove that proposed feature extraction and classifier help to improve Malaysian speaker and accent recognition.

Table 3

Accuracy of speaker recognition using English digits and Malay words

Features Extraction Method (no of coeff)		Accuracy (%) ±SD Speaker (Digits)			Accuracy (%) ±SD Speaker (Malay words)		
		KNN	SVM	ELM	KNN	SVM	ELM
MFCC (13)		88.44 ±0.12	91.49 ±0.12	89.57 ±0.09	89.91 ±0.13	92.53 ±0.12	91.24 ±0.14
LPC (16)		86.69 ±0.13	90.41 ±0.16	87.89 ±0.16	87.82 ±0.09	92.21 ±0.18	89.85 ±0.12
DWPT	Energy Entropy (62)	84.08 ±0.14	90.49 ±0.17	90.79 ±0.15	84.55 ±0.21	92.35 ±0.14	91.70 ±0.14
	Renyi Entropy (62)	83.91 ±0.17	90.79 ±0.17	90.74 ±0.13	84.49 ±0.20	92.61 ±0.17	91.50 ±0.08
	Shannon Entropy (62)	79.64 ±0.22	89.11 ±0.18	86.46 ±0.18	80.59 ±0.19	90.69 ±0.12	88.36 ±0.16

Table 3 *(Continue)*

Features Extraction Method (no of coeff)		Accuracy (%) ±SD Speaker (Digits)			Accuracy (%) ±SD Speaker (Malay words)		
		KNN	SVM	ELM	KNN	SVM	ELM
DT-CWPT	Energy Entropy (124)	84.51 ±0.22	90.86 ±0.11	92.13 ±0.11	85.03 ±0.17	93.06 ±0.18	**93.54 ±0.13**
	Renyi Entropy (124)	84.38 ±0.13	91.24 ±0.15	**92.16 ±0.12**	85.18 ±0.10	93.07 ±0.17	93.49 ±0.12
	Shannon Entropy (124)	80.37 ±0.17	90.34 ±0.23	89.39 ±0.20	81.89 ±0.14	92.00 ±0.22	86.94 ±0.17
WPT	Energy Entropy (62)	80.12 ±0.26	88.07 ±0.16	87.73 ±0.19	79.92 ±0.21	89.44 ±0.17	88.52 ±0.12
	Renyi Entropy (62)	80.21 ±0.16	87.79 ±0.10	87.66 ±0.10	79.56 ±0.18	89.44 ±0.18	87.92 ±0.16
	Shannon Entropy (62)	73.61 ±0.32	86.15 ±0.21	82.13 ±0.12	75.06 ±0.10	87.75 ±0.21	84.08 ±0.09

Table 4

Accuracy of accent recognition using English digits and Malay words

Features Extraction Method		Accuracy (%) ± SD Accent (Digits)			Accuracy (%) ± SD Accent (Malay words)		
		KNN	SVM	ELM	KNN	SVM	ELM
MFCC (13)		93.30 ±0.17	**94.48 ±0.14**	94.38 ±0.11	94.76 ±0.12	**95.86 ±0.14**	95.50 ±0.09
LPC (16)		92.05 ±0.10	93.56 ±0.08	92.81 ±0.10	92.74 ±0.14	94.73 ±0.13	93.59 ±0.08
DWPT	Energy Entropy (62)	90.94 ±0.17	93.61 ±0.17	92.52 ±0.14	91.17 ±0.18	94.80 ±0.10	94.00 ±0.15
	Renyi Entropy (62)	90.82 ±0.11	93.85 ±0.18	92.87 ±0.21	91.26 ±0.12	94.73 ±0.10	93.90 ±0.14
	Shannon Entropy (62)	87.42 ±0.16	91.03 ±0.22	87.90 ±0.23	88.09 ±0.18	92.19 ±0.12	89.92 ±0.17

Table 4 *(Continue)*

Features Extraction Method		Accuracy (%) ± SD Accent (Digits)			Accuracy (%) ± SD Accent (Malay words)		
		KNN	SVM	ELM	KNN	SVM	ELM
DT- CWPT	Energy Entropy (124)	91.12 ±0.17	93.75 ±0.14	93.60 ±0.14	91.39 ±0.18	94.61 ±0.15	91.62 ±0.21
	Renyi Entropy (124)	90.74 ±0.13	94.03 ±0.11	94.09 ±0.13	91.60 ±0.17	95.09 ±0.11	95.21 ±0.12
	Shannon Entropy (124)	87.94 ±0.21	91.84 ±0.20	90.38 ±0.17	89.22 ±0.13	92.95 ±0.12	92.14 ±0.13
WPT	Energy Entropy (62)	88.76 ±0.14	91.52 ±0.19	90.26 ±0.13	88.01 ±0.17	91.88 ±0.13	91.1 ±0.13
	Renyi Entropy (62)	88.73 ±0.21	91.57 ±0.15	90.07 ±0.17	88.09 ±0.23	92.00 ±0.10	91.01 ±0.17
	Shannon Entropy (62)	83.86 ±0.18	88.71 ±0.27	84.63 ±0.19	85.23 ±0.18	89.66 ±0.18	86.22 ±0.18

Table 5

Accuracy of speaker recognition English digits and Malay words with combination features

Features Extraction Method (no of coeff)	Accuracy (%) ± SD Speaker (digits)			Accuracy (%) ± SD Speaker (Malay words)		
	KNN	SVM	ELM	KNN	SVM	ELM
MFCC + LPC + DWPT + DT-CWPT + WPT-(773)	89.18 ±0.18	95.72 ±0.15	98.09 ±0.07	90.06 ±0.10	96.73 ±0.10	**98.68 ±0.08**

Table 6

Accuracy of accent recognition English digits and Malay words with combination features

Features Extraction Method (no of coeff)	Accuracy (%) ± SD Accent (digits)			Accuracy (%) ± SD Accent (Malay words)		
	KNN	SVM	ELM	KNN	SVM	ELM
MFCC + LPC + DWPT + DT-CWPT + WPT-(773)	94.02 ±0.15	96.85 ±0.16	98.15 ±0.06	94.78 ±0.12	97.85 ±0.08	**98.75 ±0.06**

CONCLUSION

This paper investigates the use of MFCC, LPC, WPT and DT- CWPT (first wavelet packet FB and second wavelet packet FB) based feature for speaker and accent recognition. SVM, KNN and ELM were used to measure the effectiveness of recognition of speaker and accent to identify speaker and accent. The accuracy is calculated for individual and combined features.

The result of this work shows best performance on accent recognition with 96.08 % for single feature extraction method and 98.98 % for the combination of the method. For speaker recognition, best performance achieved is 93.54 % for single feature extraction method and 98.92 % for the combination of features.

The result of feature extraction clearly outperforms the previous works even though we were using a new bigger features database. It is also found that accent identification gives better result for single feature extraction and combined features compared to speaker identification. The results of this study can be extended by using a bigger database with polysyllabic in Malay words to improve the Malaysian speaker and accent recognition. In the future work, feature reduction algorithms such as Principal Components Analysis (PCA), will be applied to reduce feature dimension. This will be developed to improve the results. It would be interesting to include experimenting, with different numbers of coefficients and other wavelet families, to observe the recognition result.

The study has many potential and useful in applications such as access control to computers, smart mobile attendance system, telephone banking, electronic commerce and forensic.

ACKNOWLEDGMENT

The authors would like to acknowledge the support and appreciation to Dr. Noriha Basir, Senior Lecturer of Center for International Languages (CIL), Universiti Malaysia Perlis (UniMAP) as language consultant for the wordlist used in the study.

REFERENCES

Adam, T. B., Salam, M. S., & Gunawan, T. S. (2013). Wavelet cesptral coefficients for isolated speech recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, *11*(5), 2731-2738.

Akusok, A., Björk, K. M., Miche, Y., & Lendasse, A. (2015). High-performance extreme learning machines: a complete toolbox for big data applications. *IEEE Access*, *3*, 1011-1025.

Almaadeed, N., Aggoun, A., & Amira, A. (2015).Speaker identification using multimodal neural networks and wavelet analysis. *IET Biometrics*, *4*(1), 18-28. doi: https://doi.org/10.1049/iet-bmt.2014.0011

Amami, R., Ayed, D. B., & Ellouze, N. (2015). Practical selection of svm supervised parameters with different feature representations for vowel recognition. *International Journal of Digital Content Technology and Its Applications(JDCTA)*, *7*(9), 418-424. doi: https://doi.org/10.4156/jdcta.vol7.issue9.50

Anand, R., Singh, J., Tiwari, M., Jains, V., & Rathore, S. (2012). Biometrics security technology with speaker recognition. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, *1*(10), 232-236.

Bong, S. Z., Wan, K., Murugappan, M., Ibrahim, N. M., Rajamanickam, Y., & Mohamad, K. (2017). Implementation of wavelet packet transform and non linear analysis for emotion classification in stroke patient using brain signals. *Biomedical Signal Processing and Control*, *36*, 102-112. doi: https://doi.org/10.1016/j.bspc.2017.03.016

Cao, J., Yang, J., Wang, Y., Wang, D., & Shi, Y. (2015). Extreme learning machine for reservoir parameter estimation in heterogeneous sandstone reservoir. *Mathematical Problems in Engineering*, *2015*, 1-10.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1-27. doi: https://doi.org/10.1145/1961189.1961199

Chelali, F. Z., & Djeradi, A. (2017). Text dependant speaker recognition using MFCC, LPC and DWT. *International Journal of Speech Technology*, *20*(3), 725-740. doi: https://doi.org/10.1007/s10772-017-9441-1

Cohen, A., Daubechies, I., & Feauveau, J. C. (2006). Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *45*(5), 485-560.

Gruhn, R. E., Minker, W., & Nakamura, S. (2011). *Statistical pronunciation modeling for non-native speech processing*. Dordrecht: Springer. doi: https://doi.org/10.1007/978-3-642-19586-0

Hanifa, R. M., Isa, K., & Mohamad, S. (2017, April 24-25). Malay speech recognition for different ethnic speakers : An exploratory study. In *2017 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)* (pp. 91-96). Langkawi, Malaysia. doi: https://doi.org/10.1109/ISCAIE.2017.8074956

Hariharan, M., Chee, L. S., Ai, O. C., & Yaacob, S. (2012). Classification of speech dysfluencies using lpc based parameterization techniques. *Journal of Medical Systems*, *36*(3), 1821-1830. doi: https://doi.org/10.1007/s10916-010-9641-6

Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. *PLoS ONE*, *11*(7), 1-21. doi: https://doi.org/10.1371/journal.pone.0158520

Jain, A., & Sharma, O. P. (2013). A vector quantization approach for voice recognition using mel frequency cepstral coefficient (MFCC): a review. *International Journal of Electronics and Communication Technology*, *4*(4), 26-29.

Johari, N. A. A. B., Hariharan, M., Saidatul, A., & Yaacob, S. (2011, October 20-21). Multistyle classification of speech under stress using wavelet packet energy and entropy features. In *2011 IEEE Conference on Sustainable Utilization Development in Engineering and Technology, STUDENT 2011* (pp. 74-78). Semenyih, Malaysia. doi: https://doi.org/10.1109/STUDENT.2011.6089328

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12-40. doi: https://doi.org/10.1016/j.specom.2009.08.009

Lee, D. T., & Yamamoto, A. (1994). Wavelet analysis: theory and applications. *Hewlett Packard Journal, 45*, 44-52. doi: https://doi.org/10.1051/jp1:1997114

Lei, L., & Kun, S. (2017). Speaker Recognition Using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring. *Journal of Electrical and Computer Engineering*, *2017*, 1-9. doi: https://doi.org/10.1155/2017/1735698

Li, X., & Yu, W. (2014). Fast support vector machine classification for large data sets. *International Journal of Computational Intelligence Systems*, *7*(2), 197-212. doi: https://doi.org/10.1080/18756891.2013.868148

Liu, C. L., Lee, C. H., & Lin, P. M. (2010). A fall detection system using k-nearest neighbor classifier. *Expert Systems with Applications*, *37*(10), 7174-7181. https://doi.org/10.1016/j.eswa.2010.04.014

Lim, W. J., Muthusamy, H., Yazid, H., Yaacob, S., Nadarajaw, T., Lim, W. J., … & Yaacob, S. (2016, August 10-12). Dual tree complex wavelet packet transform based infant cry classification. In *AIP Conference Proceedings* (Vol. 1775, No. 1, p. 30049). Songkhla, Thailand. doi: https://doi.org/10.1063/1.4965169

Mohd Yusof, S. A., & Yaacob, S. (2008). Classification of Malaysian vowels using formant-based features. *Journal of ICT*, *7*, 27-40.

Oung, Q. W., Muthusamy, H., Basah, S. N., Lee, H., & Vijean, V. (2018). Empirical wavelet transform based features for classification of Parkinson's disease severity. *Journal of Medical Systems*, *42*(2), 29-45.

Pandiaraj, S., & Kumar, K. R. S. (2015). Speaker identification using discrete wavelet transform. *Journal of Computer Science*, *11*(1), 53-56. doi: https://doi.org/10.3844/jcssp.2015.53.56

Paulraj, M. P., Yaacob, S., & Mohd Yusof, S. A. (2008, July 7-9). Vowel recognition based on frequency ranges determined by bandwidth approach. In *ICALIP 2008 - 2008 International Conference on Audio, Language and Image Processing* (pp. 75-79). Shanghai, China. doi: https://doi.org/10.1109/ICALIP.2008.4590133

Prasanna, S. M., Gupta, C. S., & Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, *48*(10), 1243-1261. doi: https://doi.org/10.1016/j.specom.2006.06.002

Rathor, S., & Jadon, R. S. (2017, July 3-5). Text indpendent speaker recognition using wavelet cepstral coefficient and butter worth filter. In *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017* (pp. 1-5). Delhi, India. doi: https://doi.org/10.1109/ICCCNT.2017.8204079

Saksamudre, S. K., & Deshmukh, R. R. (2015). Comparative study of isolated word recognition system for Hindi language. International Journal of Engineering Research and Technology (IJERT), 4(07), 536-540. doi: https://doi.org/10.17577/IJERTV4IS070443

Sangeetha, S., & Radha, N. (2013, January 4-5). A new framework for IRIS and fingerprint recognition using SVM classification and extreme learning machine based on score level fusion. In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on* (pp. 183-188). Coimbatore, India.

Soon, M. W., Anuar, M. I. H., Abidin, M. H. Z., Azaman, A. S., & Noor, N. M. (2017, September 12-14). Speech recognition using facial sEMG. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 1-5). Kuching, Malaysia. doi: https://doi.org/10.1109/ICSIPA.2017.8120569

Valyon, J., & Horváth, G. (2003). A weighted generalized LS-SVM. *Periodica Polytechnica Electrical Engineering*, *47*(3-4), 229-251.

Yadav, S. S., & Bhalke, D. G. (2015). Speaker identification system using wavelet transform and VQ modeling technique. *International Journal of Computer Applications*, *112*(9), 19-23.

Yusnita, M. A., Paulraj, M. P., Yaacob, S., Bakar, S. A., & Saidatul, A. (2011, November 25-27). Malaysian English accents identification using lpc and formant analysis. In *Proceedings - 2011 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2011* (pp. 472-476). Penang, Malaysia. doi: https://doi.org/10.1109/ICCSCE.2011.6190572

Yusnita, M. A., Paulraj, M. P., Yaacob, S., & Shahriman, A. B. (2012, December 3-4). Classification of speaker accent using hybrid dwt-lpc feature and k-nearest neighbors in Malaysian English accented speech. In *2012 IEEE Symposium on Computer Applications & Industrial Electronics* (pp. 179-184). Kota Kinabalu, Malaysia.

Zhang, Y., Dong, Z., Wang, S., Ji, G., & Yang, J. (2015). Preclinical diagnosis of magnetic resonance (MR) brain images via discrete wavelet packet transform with tsallis entropy and generalized eigenvalue proximal support vector machine (GEPSVM). *Entropy*, *17*(4), 1795-1813. doi: https://doi.org/10.3390/e17041795