Feature Reduction Method for Speaker Identification Systems Using Particle Swarm Optimization

Ahmed Al-Hmouz [#1], Khaled Daqrouq [*2], Rami Al-Hmouz [*2], Jaafar Alghazo [#3]

[#1]Department of Information Technology, Middle East University, Jordan.
[1] aa998@uowmail.edu.au

[*2]Department of Electrical and Computer Engineering, King Abdulaziz University, Saudi Arabia.
[2]haleddaq@yahoo.com
[2]ralhmouz@kau.edu.sa

[#3]College of Computer Engineering and Science Computer Engineering,
Prince Mohammad bin Fahad University, Saudi Arabia.
[3] jghazo@pmu.edu.sa

***Abstract*—Feature selection (FS) is a process in which the most informative and descriptive characteristics of a signal that will lead to better classification are chosen. The process is utilized in many areas, such as machine learning, pattern recognition and signal processing. FS reduces the dimensionality of a signal and preserves the most informative features for further processing. A speech signal can consist of thousands of features. Feature extraction methods such as Average Framing Linear Prediction Coding (AFLPC) using wavelet transform reduce the number of features from thousands to hundreds. However, the vector of features involves some redundancy. In addition, some features are similar and do not give discrimination to classes. Taking such features into consideration in the classification process will not help to identify certain classes; conversely, they will only serve to confuse the classifier and inhibit identification of accurate classes. This paper proposes an FS method that uses evolution optimization techniques to select the most informative features that maximize the classification rates of Bayesian classifiers. The classification rate is also maximized by modeling the features with the proper number of Gaussian distributions. The results of comparative analysis conducted show that the selection based individual speaker model gives the best classification rate performance.**

Keyword - Feature Selection, Speaker Identification, Bayes Theorem.

I. INTRODUCTION

Research on automatic speech recognition (ASR) has actively been conductedover the past four decades[1]. ASR is a tool with many potential applications such as automation of operator-assisted services and speech to text systems for hearing-impaired individuals [2]. In speaker recognition systems, the speech signal is represented by several features, which play a major part in system design. Karhunen-Loeve transform (KLT) based features [3], Mel Frequency Cepstral Coefficient (MFCC) [4], Linear Predictive Cepstral Coefficient (LPCC) [5], and wavelet transform-based features [6]-[8] are examples of signal speech features.

Various approaches have been proposed to reduce the number of features required for speech recognition. Paliwal[9] reduced the dimensionality of feature vectors in speech recognition systems and tested the technique on four methods. In [10], the Laplacian Eigenmaps Latent Variable Model (LEVLM) used fewer MFCC vectors without affecting the recognition rate, and it exhibited better performance than Principal Component Analysis (PCA). Feature frame selection based on phonetic information has also been investigated to increase classification rate; however, the exact phonemes cannot be easily extracted [11].

Joint factor analysis (JFA) [12]-[14]is commonly used to enhance the performance of text independent speaker verification systems by modeling speaker and session variability. This work has been extended to i-vector, which outperforms JFA in terms of complexity and model size [15]. The classification rate increaseswith the number of feature frames available for training and testing [16]. However, the performance does not continue to improve if more features are added and redundancies exist in the features; consequently, some features can be ignored with no effect on recognition performance [17].

Researchers have also focused on selecting valuable features in speech recognition systems. For example, Euclidean distance measure has been used to determine frame rate [18], anentropy-based approach has been utilized in speech signal in-frame selection [19], andmaximum likelihood-based feature selection has also been investigated [20]. The previous methods select valuable features in speech signals, but they tend to ignore the redundancy of features in feature frames. Further, reports indicate that maximizing feature information does not lead to a better classification rate [21], [22]. Features should contain minimum redundancy within the selected

features in the speaker model [11]. Therefore, valuable features as well as some redundancy should be available in the selection to improve performance. In this context, searching for such features requires a heuristic random search and evolutionary computation techniques.

In general, the Gaussian Mixture Model [23] is used to model features in speaker models for speaker identification, usually by means of 5–10 Gaussian densities. In this method, all features are forced to have a specific number of Gaussians. However, the classification rates vary according to the number and type of Gaussian. An iterative process for choosing the final estimation with the highest likelihood [24] has also been investigated. Various other methods have also been proposed [25]-[27] to estimate the optimal number of mixtures; however, they estimate the optimal number of mixtures only between two values (minimum and maximum). According to the nature of features, some features require higher/lower Gaussian densities than other features for accurate modeling.

Feature extraction methods transform the speech signal from the time domain into another feature space domain, with the coefficients in the new domain being less than the frame size of speech signal. The process of transformation produces a set of features representing the speech signal; however, some features produced in the speaker models do not give effective distinction to classes. This paper proposes a method that improves classifier performance by removing and preserving redundancy in features within the same class and among all classes in speaker identification systems, regardless of the feature extraction method employed. The proposed feature selection method is based on the available features that maximize the classification rate. The selection can be achieved individually (each class has its own set of features) or globally (classes settle on a set of featuresas a group).

Determining the effective set of features is achieved by using the Particle Swarm Optimization (PSO) evolutionary optimization technique. We also use PSO to determine the optimal number of Gaussian densities in order to model each feature in the feature vector space that leads to a better classification rate when coupled with wavelet-based feature extraction methods[28] and the Bayes classifier. PSO was proposed in [29], [30] for feature selection in speaker verification systems using a binary classification process. In that work, the selection was considered from different aspects of the speaker identification system, and the selection process was realized on all speaker models, and the best featuremodels chosen.

The length of the feature vector at the input of the classifier is crucial in the classification process; these features contribute the most to recognition rate in the classification. There is no gain to consider extra feature in the classification process unless they are informative. In this paper, we propose a method that selects the most informative features for a given feature vector. Two selection approaches are presented: first, all classes settle on a group of features that maximize classification rate; second, each speaker model selects its own set of features that are different from other speakers. Further, features are modeled as one or two Gaussian densities considering extraction of features using AFLPC [28]. Consequently, selection of the exact number of Gaussian densities that maximize classification rate is also considered in this paper.

The remainder of this paper is organized as follows.Section IIdiscussesthe wavelet-based feature AFLPC. Section III outlines the proposed feature selection method. Section IVpresents the experimental results obtained.Finally, Section V concludes this paper.

II. THE AFLPC FEATURE EXTRACTION METHOD

Wavelet packets can be used to extract additional features to guarantee a higher recognition rate. Avciet al. [31] proposed a method that calculates the entropy value of the wavelet norm in digital modulation recognition. A robust speech recognition scheme that uses wavelet-based energy as a threshold for denoising estimation has also been proposed for noisy environments [32]. Wu and Lin [33] proposed a method that uses the energy indexes of Wavelet Packet (WP) for speaker identification. Entropy calculation for the waveforms at the terminal node signals obtained from Discrete Wavelet Transform (DWT) has also been used in speaker identification [34].

Avci[35] investigated a feature extraction method for speaker recognition based on a combination of three entropy types (sure, logarithmic energy, and norm) was investigated. Daqrouq and Al Azzawi[28] and Wu and Lin [36] also proposed using DWT w instead of the Discrete Cosine Transform (DCT) to solve the problem of high frequency artifacts being introduced as a result of abrupt changes at window boundaries. The features based on DWT were chosen to evaluate the effectiveness of the selected feature for speaker identification [28], [37]. Several levels of DWT approximation sub-signals exhibited good performance in the presence of Additive White Gaussian Noise (AWGN) [37].

Before the feature extraction stage, the speech data are processed by a silence-removing algorithm followed by application of a pre-processing technique. In AFLPC, features are extracted from the$Z$frames of each WT speech sub-signal:

$$\{U_q(t)\} = \{u_{q1}(t), u_{q2}(t) \dots . u_{qZ}(t)\} \quad (1)$$

where$Z$ is the number of considered frames (each frame of 20ms duration) for theq^{th} WT sub-signal$u_q(t)$ and $t$is the discrete time. The average of the LPC coefficients calculated for the$Z$frames of$u_q(t)$is utilized to extract a wavelet sub-signal feature vector as follows:

$$f_q = \sum_{z=1}^{Z} LPC(u_{qz}(t))\frac{1}{Z} \quad (2)$$

The feature vector of the entire given speech signal is

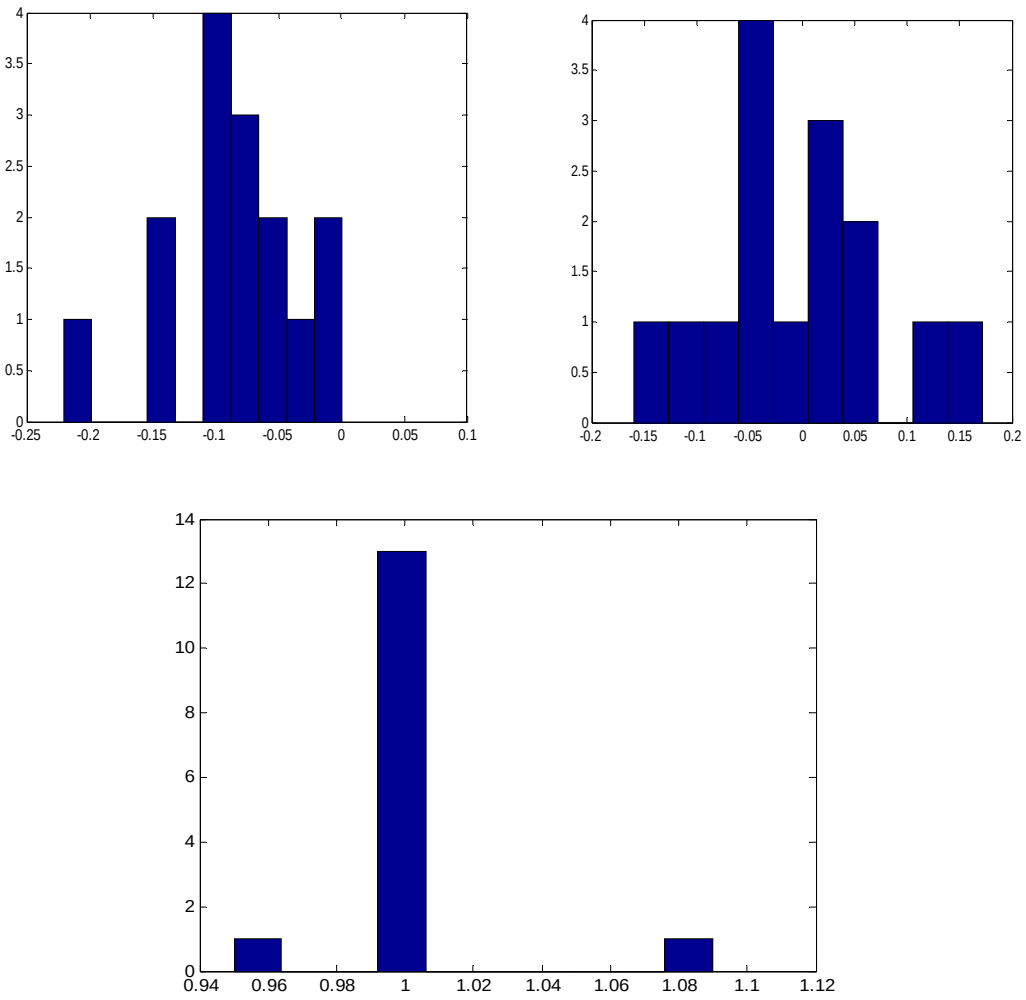
$$AFLPC = \{f_1, f_2, \ldots\ldots, f_Q\} \quad (3)$$

In this paper, the combination of AFLPC and WP is denoted WPAFL and that of AFLPC and DWT is denoted DWTAFL.

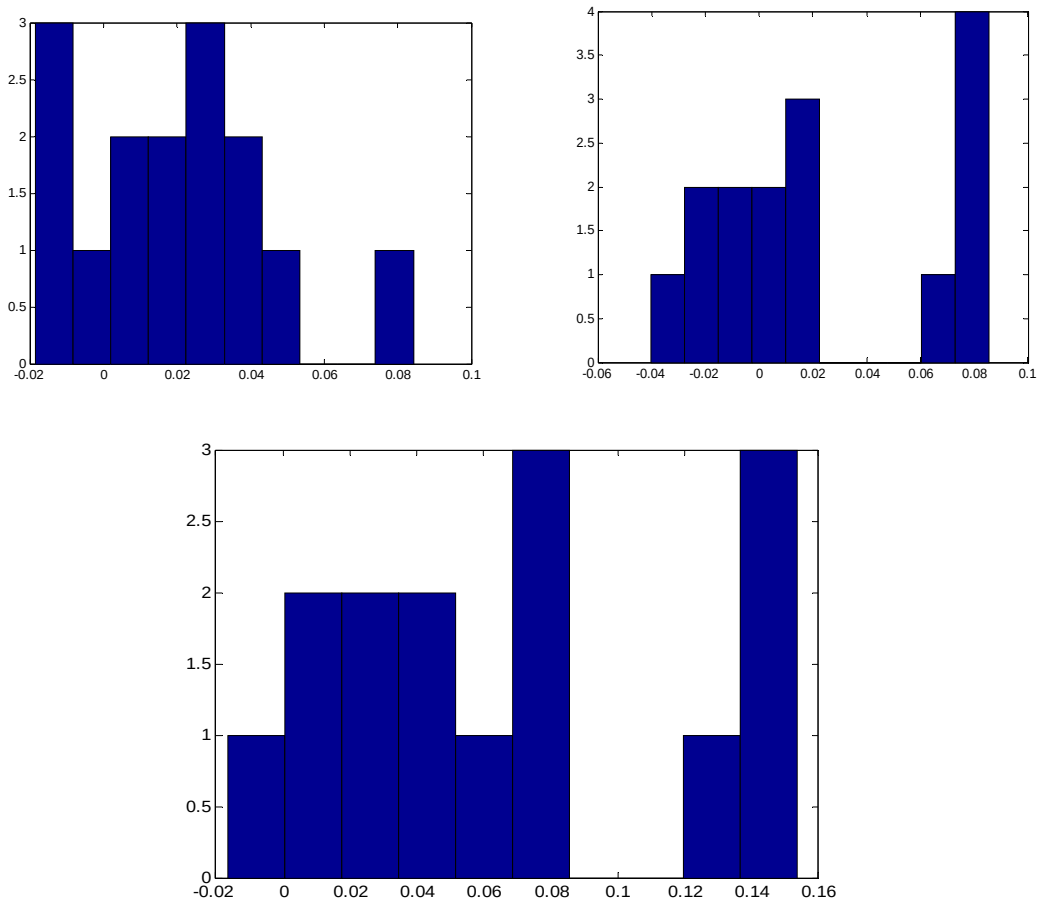
III. Feature Selection

A. *Feature Modeling*

In AFLPC, features are extracted from the speech signalbased on wavelet transforms. A naïve Bayes classifier can then be used to perform recognition, and the distribution of features can be modeled as a Gaussian—the distributions of some selected features are shown in Fig. 1(a). However, some features can also be modeled as two Gaussians, as shown in Fig. 1(b). These two models are sufficient for realization of classification after determining the indices of the features in order to consider the exact model, as will be shown in the results obtained.



(a)



(b)

Fig. 1: Feature (AFLPC) distributions of speech signals (x-axis: feature value, y-axis; repetition):
(a) Features can be modeled with one Gaussian. (b) Features can be modeled with more than one Gaussian.

In AFLPC, some features should be modeled as one Gaussian and other features with more than one. The choice is crucial for building likelihood functions as better representation of features leads to better classification rates. For the Bayesian fusion process, let$f_1, f_2 \dots f_Q$be features that have been produced from AFLPC, and$C_1, C_2 \dots C_M$be the available classes. The probability of$P(C_m/f_1, f_2 \dots f_Q)$is calculated using Bayes rule:

$$P(\frac{C_m}{f_1 f_2} \dots f_Q) = \frac{P(f_1 f_2 \dots\dots f_N) P(C_m)}{\sum_{j=1}^{M} P(f_1 f_2 \dots\dots f_Q / C_j) P(C_j)} \quad (4)$$

where Q is the total number of features in AFLPC, M is the total number of available classes, and$P(f_1, f_2 \dots f_Q / C_m)$is the likelihood function. Surprisingly, the Naive Bayes model performs well, even in situations where independence assumptions are clearly false [38]. Using the assumption of conditional independence to reduce the number of parameters, we get

$$P(C_m/f_1, f_2 \dots f_Q) = \frac{\prod_{k=1}^{Q} P(f_k/C_m)\, P(C_m)}{\sum_{j=1}^{M} P(f_1, f_2 \dots f_Q)\, P(C_j)} \quad (5)$$

The Posterior$P(C_m/f_1, f_2 \dots f_Q)$is computed by multiplying the feature probabilities in the speaker signal. Features are modeled as one Gaussian or a mixture of Gaussians (two Gaussians) according to the nature of the feature.$P(C_m)$is the prior probability of class m, it is assumed that all classes are equally likely$P(C) = 1/m$. $\sum_{j=1}^{M} P(f_1 f_2 \dots\dots f_Q / C_j) P(C_j)$ is a normalization term. The maximum a posterior probability (MAP) is used to estimate the speaker class Ci that maximizes$P(C_i/f_1, f_2 \dots f_Q)$:

$$\arg maz_i \left\{ P\left(C_i / f_1, f_2 \dots f_Q \right) \right\} \quad (6)$$

B. Evolutionary Optimization of the Classifier

Choosing the number of Gaussians for the feature vector cannot be achieved by observing every single feature and deciding on the number of suitable Gaussians in the feature model. There should be an iterative process to settle on the most suitable number of Gaussians. With this in mind, we considered more advanced methods that focus on minimization of the classification error (maximization of classification rate). Consequently, we decided on evolutionary optimization or population-based optimization owing to the flexibility of the fitness function supported by the nature of the optimization process.

To make the fitness function fully reflective of the performance of the classifier (so that the classifier can be effectively optimized), we determine the number of Gaussian mixture distributions$j$ (one or two) in the feature vector$f_1, f_2 \dots f_Q$for a given speech signal in the classes$C_1 - C_M$, where

$$j1 - \frac{Q}{C1} - M = \arg j \max R \qquad (7)$$

The classification error is minimized or the classification accuracy $(R)$is maximized. In other words, we look for the number of Gaussians$j$in each feature, in each speech signal, in the training set that maximizes the classification rate R. (In our experiments, we use PSO owing to its simplicity, relatively low computing overhead, and high effectiveness.

In certain locations of features$f_1, f_2 \dots f_Q$, the statistics for these features are alike for all classes and some of them do not give discriminations to classes. Considering such features in the classification process will not enhance recognition decisions but instead may minimize the probability of some classes or increase the probability of other classes in which misclassification might occur.

We wish to maximize the classification R based on the available feature set. Therefore, a process that will select the most effective set of features that discriminate all classes is required. Reducing the number of features while increasing the classification rate reduces system complexity. Thus, the question is whether to consider f or not in the classification process. The number of features is minimized in all classes, but these features are considered more representative in the sense that the probability of the desired class will increase and the probability of other classes will decrease. PSO is also used for maximization problems.

$$F_s/C_{1-M} = \arg j \max R.F_s C \left\{f_1, f_2 \dots f_Q\right\} \qquad (8)$$

It should be noted that the indices of Fs are the same for all speakers. The posterior of all speakers m (Eqs. 5 and 6) will be calculated for given selected features Fs. The number of selected features nis determined by the optimization process. It can also be noted that there are distinctive features for individual speakers that are different in indices from other speakers. With this in mind, each speaker (class) can have its own selected features that maximize the classification rate for that particular class. In this case, the selection process is more complex in terms of number of features as each class has its own selected features compared to the problem in Eq. 7. To reduce the optimization complexity, the number of selected features is first predetermined in each class. Then, PSO can be executed to maximize R, with the result being a group of selected features for each class. It is not necessary for the indices of the features to be the same for all classes.

We considered three cases:

1) Features are modeled as one Gaussian: In this case, R is maximized with regards to features$f_1, f_2 \dots f_Q$. The most representative features$F_s$are considered by running PSO on the training set. Feature$f_i$is either selected and its$P(f_i/C_m)$ considered in Eq. 5 or ignored and$P(f_i/C_m)$considered as one for allC_m. In the worst case, if the selection does not produce a better classification rate, it will at least give the same performance with fewer features, which in the end affect system complexity.

2) Features are modeled as one or two Gaussians: In this case, R is maximized with regards to features$f_1, f_2 \dots f_Q$, as in case 1, and when considering$f_i$the suitable models (one or two Gaussians) is also considered in the optimization process.

3) Considering feature modeling in case 2, each class has its own selected feature set. Here, the number of features is given before running PSO. If n features of each class are assumed to be selected among$f_1, f_2 \dots f_Q$, then the number of parameters that have to be optimized is $n \times M$.

Note that considering more than two Gaussians will not improve the classification rate as most features are distributed normally and few of them have the nature of bimodal Gaussian distribution, as shown in Fig. 1.

IV. EXPERIMENTS AND RESULTS

The experimental setup was as follows: speech signals were recorded via a PC sound card at a spectral frequency of 4000 Hz and sampling frequency of 8000 Hz. Forty-seven persons participated in the recordings. Each participant recorded a minimum of 20 different utterances in Arabic. The age of the speakers ranged from 20 to 45 years and the participants comprised 25 male and 22 females. The recording process was provided in normal university office conditions. The speech signals data set was split into training set (496) and test set (530). Two speech signals in the training set for all classes were left for feature selection using PSO and the rest were used to form likelihood functions. It should be noted the presented method improve the classification rate of existing feature extraction methods to provide the informative features to the classifier.

A neural network was also used to evaluate the feature selection performance. The same optimization process was used to optimize the number of considered features at the input of the neural network. The features produced from AFLPC were fed directly into the input of a probabilistic Neural Network.

For PSO, the size of the population was set to 25 and the number of iterations was set to 100; we experimentally found that this number was sufficient to achieve the convergence of the method (Fig. 2). PSO has been shown to be robust and effective in solving the optimization problem. The values of the cognitive acceleration coefficient and the social acceleration coefficient were set to 0.5 and 1.25, respectively. These two values are commonly recommended in the literature (see [39]). PSO is used because it outperforms other heuristic search methods (i.e., Genetic algorithms) in terms of convergence speed and complexity [29].

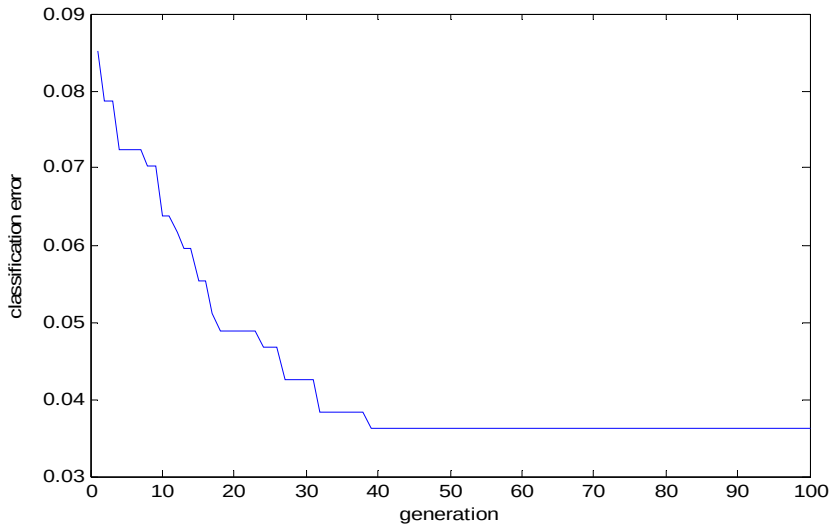


Fig 2. Sample of PSO Convergence

A. *Case 1:*

Table 1 shows the classification rates of various feature extraction methods considering two classifiers, Bayes classifier (BC) and GMM, and the probabilistic neural network (PNN), along with the number of features obtained for PSO. The performance of all the feature extraction methods improved using feature selection by considering fewer features. In our evaluation of the proposed method, several published methods were analysed.

Feature extraction such as Wavelet packet and Shannon entropy (WPS) [37], wavelet packet and Log energy entropy (WPLE) [35], MFCC and GMM (MFGMM) [23], WPAFL, DWAFL, as well as the fusion between WPAFL and DWAFL (FWAFL) were used to test the proposed method. In general, there were improvements in the classification rates regardless of the feature extraction method and the type of classifier. AFLPC can be reduced by as much as 50%, with at least the same performance in terms of classification rate. The choice of classifier has no effect as feature selection improves the classification rate for NN. The speaker identification system is consequently less complex as the number of features at the input of the NN is approximately 50%. MFCC with GMM has the best performance with classification rate reaches 0.9815.

B. *Case 2:*

Table 2 shows the feature selection performance achieved for one Gaussian, two Gaussians, and a group comprising one and two Gaussians. The optimization process determines the number of Gaussians that should be considered in the realization. As in case 1, the feature reduction is as much as 50% for the AFLPC methods, with improvements of 3.04% and 1.55% for DWTAFL and WPAFL, respectively. Note that the more features

of the AFLPC (FWAFL) method there are, the better the performance with respect to classification rates, more information is provided to classes.

TABLE 1. Feature Reduction and Classification Rates for Case 1 (no PSO: no optimization).

Feature extraction Method	Number of features	Train no PSO	Train PSO	Test no PSO	Test PSO	Number of selected Features (percentage, %)	Classifier	Improvement (%)*
DWTAFL	186	0.9906	1	0.8810	0.9254	143 (76.9%)	BC (one Gaussian)	5.04%
WPAFL	186	0.9662	0.9919	0.9093	0.9375	96 (50.6%)	BC (one Gaussian)	3.10
FWAFL	372	0.9700	1	0.9355	0.9758	184 (49.4%)	BC (one Gaussian)	4.31%
MFGMM	42	0.9962	0.9962	0.9733	0.9815	24 (57.1%)	BC (5 Gaussians)	0.84%
WPS	127	0.9662	0.9457	0.6039	0.6334	56 (44.1%)	PNN	4.88%
WPLE	64	0 .9962	0.9962	0.5282	0.5338	40 (62.5%)	PNN	1.06%
WPAFL	930	1	0.9944	0.9254	0.9274	518 (55.7%)	PNN	0.21%

* (Test PSO - Test no PSO)/Test no PSO

TABLE 2. Feature Reduction and Classification Rates for Case 2.

Feature extraction method	No of features	Train no PSO		Train PSO (Mix) features: (# GMM)	Test no PSO		Test PSO (Mix) No selection	Test PSO Feature Selection	No. of selected Features (%)	Improvement (%) *
		1GMM	2GMM		1GMM	2GMM				
DWTAFL	186	0.9906	0.9906	0.9905 92: (1GMM) 94: (2 GMM)	0.8810	0.8156	0.9254	0.9274	145 (77.8%)	5.27%
WPAFL	186	0.9662	0.9662	0.9662 92: (1GMM) 94: (2 GMM)	0.9093	0.9073	0.9375	0.9395	104 (55.9%)	3.32%
FWAFL	372	0.9700	0.9700	0.97000 172: (1 GMM) 200: (2 GMM	0.9355	.9355	0.9758	0.9778	193 (51.8%)	4.52%

* (Test PSO Feature Selection - Test no PSO (1 GMM))/ Test no PSO (1 GMM).

For feature selection in this situation, we not only get better performance but also a reduction in the number of features used in the classification process. It was found that modeling the features with more than two Gaussians will not improve the preface of classification rate as the best models of most features will be modeling with only one Gaussian.

C. Case 3:

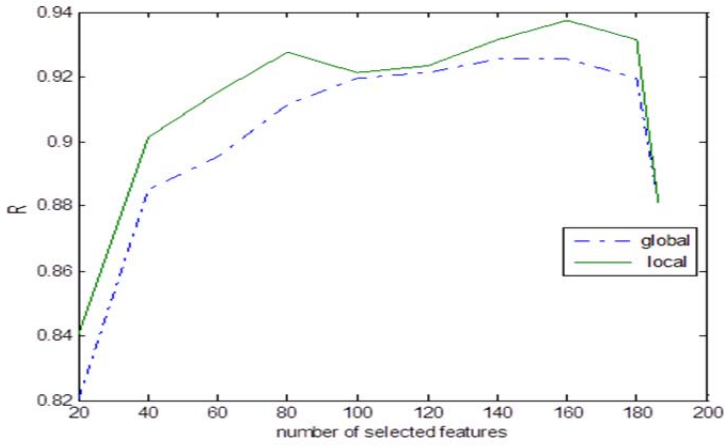
In this case, the number of parameters (selected features) was increased as each class had n number of selected features. In another words, each class settled on a set of features that discriminate it the most from other classes; the selected features in each class may overlap with other selected features in other classes (overlap means that the indices of the selected feature could be the same). Table 3 shows the performance of the AFLPC methods at n = 160 for DWTAFL and WPAFL and n = 180 for FWAFL, the performance is improved when each class selects its own set of features (local) in contrast to all classes settling on a set of features (global). Allowing the class to choose its selected features offer advantages over features that are selected from all classes, as can be seen in Fig. (3). It should be noted that the solution of the global feature is considered to be the initial population for the local features. Note also that there will be slight improvements in R after n = 80 for DWTAFL and WPAFL and after n = 100 for FWAFL. The complexity is increased as the number of features in the optimization process is increased (47× n).

I PSO does not guarantee an optimal solution, but the results show that there groups of features that improve the classification rate when they are selected exist regardless of the feature extraction method and the type of classifier used. The best performance was 98.59% for FWAFL in local feature selection at n = 180 for each class, with a total number of features at 8460. Genetic algorithm (GA) is also tested for FWAFL, the best performance was 97.98% when n = 240 (11280).

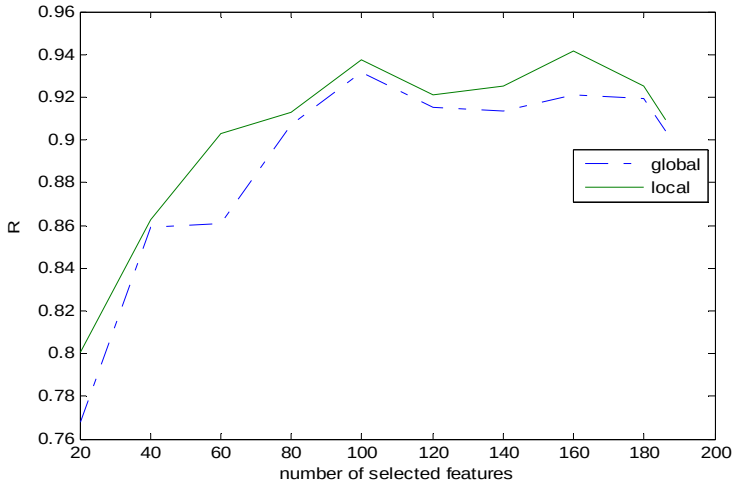
TABLE 3. Feature Reduction and Classification Rates for case 3 at n = 160 (DWTAFL, WPAFL) and 180 (FWAFL).

Feature extraction method	Number of features	Train no PSO	Train PSO (local)	Test no PSO	Test PSO (global)	Test PSO (local)	Number of selected n (global) n × M (local)	Improvement %*
DWTAFL	186	0.9906	0.9887	0.8810	0.9254	0.9375	160 (global) 7520 (local)	6.41%
WPAFL	186	0.9662	0.9587	0.9093	0.9214	0.9415	160 (global) 7520 (local)	3.54%
FWAFL	372	0.9700	0.9887	0.9355	0.9758	0.9859	180 (global) 8460 (local)	5.34%

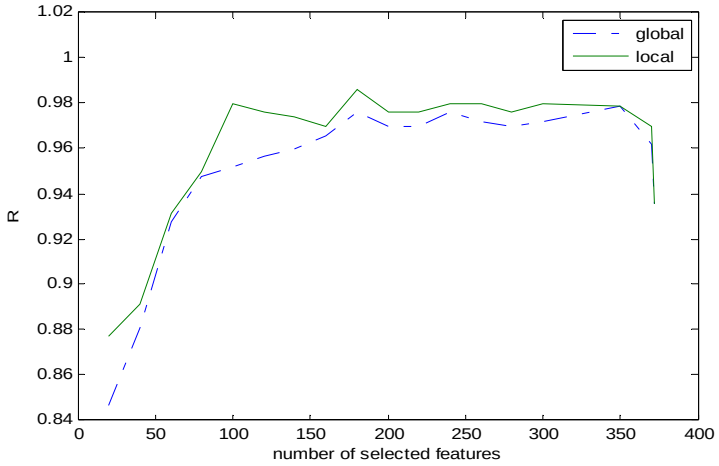
* (Test PSO Feature Selection (local) - Test no PSO)/ Test no PSO



(a)



(b)



(c)

Fig. 3: number of selected features vs classification rates: (a) DWTAFL, (b) WPAFL (c) FWAFL

V. CONCLUSION

In this paper, features that are more informative among classes were selected and considered in the classification process. Selection of features resulted from maximizing the classification rate. An evolutionary optimization technique (PSO) was used in the selection process. Feature selection guarantees at least the same performance with fewer features, especially when there is redundant information in the features. Feature extraction methods such as AFLPC produces features with redundant information. Further, some features are uninformative, in that they do not enhance the classification rate. Consequently, ignoring such features enhanced the performance of the Bayes classifier. Bayes classifier is sensitive when modeling features, but choosing the right number of Gaussian models in the feature selection, eventually improved its performance. In AFLPC, a 50% feature reduction rate was achieved with no impact on the classification rate. Each class was also identified by its own set of features. The best classification rate was 98.59% for FWAFL in local feature selection.

ACKNOWLEDGMENT

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. The authors, therefore, acknowledge with thanks DSR technical and financial support

REFERENCES

[1] Steve Young, “A review of large-vocabulary continuous-speech,” IEEE Signal Process. Mag., vol. 13, no. 5, p. 45, 1996.
[2] R. Stuckless, Developments in real-time speech-to-text communication for people with impaired hearing. … access for people with hearing loss, 1994.
[3] S.-Y. Lung and C.-C. T. Chen, “Further reduced form of Karhunen-Loeve transform for text independent speaker recognition,” Electronics Letters, vol. 34, no. 14, pp. 1380–1382, 1998.
[4] D. Hosseinzadeh and S. Krishnan, “Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs.,” MMSP, pp. 365–368, 2007.
[5] M. Afify and O. Siohan, “Comments on Vocal Tract Length Normalization Equals Linear Transformation in Cepstral Space.,” TASLP, vol. 15, no. 5, pp. 1731–1732, 2007.
[6] O. Rioul and M. Vetterli, “Wavelets and signal processing,” IEEE Signal Process. Mag., vol. 8, no. 4, pp. 14–38, 1991.
[7] C. J. Long and S. Datta, “Wavelet based feature extraction for phoneme recognition.,” ICSLP 1996, 1996.
[8] S.-Y. Lung, “Improved wavelet feature extraction using kernel analysis for text independent speaker recognition.,” DSP, vol. 20, no. 5, pp. 1400–1407, 2010.
[9] K. K. Paliwal, “Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer.,” DSP, vol. 2, no. 3, pp. 157–173, 1992.
[10] A. Jafari and F. Almasganj, “Using Laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy.,” SPEECH, vol. 52, no. 9, pp. 725–735, 2010.
[11] C.-S. Jung, M. Y. Kim, and H.-G. Kang, “Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information.,” TASLP, vol. 18, no. 6, pp. 1332–1340, 2010.
[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint Factor Analysis Versus Eigenchannels in Speaker Recognition.,” TASLP, vol. 15, no. 4, pp. 1435–1447, 2007.
[13] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and Session Variability in GMM-Based Speaker Verification.,” TASLP, vol. 15, no. 4, pp. 1448–1460, 2007.
[14] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A Study of Interspeaker Variability in Speaker Verification.,” TASLP, vol. 16, no. 5, pp. 980–988, 2008.
[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification.,” TASLP, vol. 19, no. 4, pp. 788–798, 2011.
[16] H. Gish and M. Schmidt, “Text-independent speaker identification,” IEEE Signal Process. Mag., vol. 11, no. 4, pp. 18–32, 1994.
[17] E. Choi, “On reducing complexity of acoustic models for Mandarin speech recognition,” presented at the 9th Australian International Conference on Speech Science Technology, Melbourne, 2002, pp. 166–171.

[18] Qifeng Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," presented at the 2000 International Conference on Acoustics, Speech and Signal Processing, 2000, vol. 3, pp. 1783–1786.
[19] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," presented at the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, vol. 1, pp. I–549–52.
[20] Tingyao Wu, D. Van Compernolle, J. Duchateau, and H. Van Hamme, "Maximum Likelihood Based Temporal Frame Selection," presented at the 2006 IEEE International Conference on Acoustics Speed and Signal Processing, 2006, vol. 1, pp. I–349–I–352.
[21] D. P. W. Ellis and J. A. Bilmes, "Using mutual information to design feature combinations.," INTERSPEECH, pp. 79–82, 2000.
[22] T. Eriksson, S. Kim, Hong-Goo Kang, and Chungyong Lee, "An information-theoretic perspective on feature selection in speaker recognition," IEEE Signal Processing Letters, vol. 12, no. 7, pp. 500–503, 2005.
[23] 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.
[24] S. Richardson and P. Green, Bayesian Approaches to Gaussian Mixture Modeling. IEEE Trans. on PAMI, 1997.
[25] Zheng Rong Yang and M. Zwolinski, "Mutual information theory for adaptive mixture models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 4, pp. 396–403, Apr. 2001.
[26] A. Likas, N. Vlassis, and J. J Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 36, no. 2, pp. 451–461, Feb. 2003.
[27] J. J. Verbeek, N. Vlassis, and B. Kröse, "Efficient Greedy Learning of Gaussian Mixture Models," Neural Computation, vol. 15, no. 2, pp. 469–485, Feb. 2003.
[28] K. Daqrouq and K. Y. Al Azzawi, "Average framing linear prediction coding with wavelet transform for text-independent speaker identification system," Computers & Electrical Engineering, vol. 38, no. 6, pp. 1467–1479, Nov. 2012.
[29] R. Hassan, B. Cohanim, O. de Weck, and G. Venter, "A Comparison of Particle Swarm Optimization and the Genetic Algorithm," presented at the 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Reston, Virigina, 2012.
[30] S. Nemati and M. E. Basiri, "Particle Swarm Optimization for Feature Selection in Speaker Verification," in Applications of Evolutionary Computation, vol. 6024, no. 39, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 371–380.
[31] E. AVCI, D. HANBAY, and A. VAROL, "An expert Discrete Wavelet Adaptive Network Based Fuzzy Inference System for digital modulation recognition," Expert Systems with Applications, vol. 33, no. 3, pp. 582–589, Oct. 2007.
[32] W. Lu, W. Sun, and H. Lu, "Robust watermarking based on DWT and nonnegative matrix factorization," Computers & Electrical Engineering, vol. 35, no. 1, pp. 183–188, Jan. 2009.
[33] J.-D. Wu and B.-F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," Expert Systems with Applications, vol. 36, no. 2, pp. 3136–3143, Mar. 2009.
[34] D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy," Expert Systems with Applications, vol. 36, no. 3, pp. 6295–6300, Apr. 2009.
[35] E. Avci, "A new optimum feature extraction and classification method for speaker recognition: GWPNN," Expert Systems with Applications, vol. 32, no. 2, pp. 485–498, Feb. 2007.
[36] J.-D. Wu and B.-F. Lin, "Speaker identification based on the frame linear predictive coding spectrum technique," Expert Systems with Applications, vol. 36, no. 4, pp. 8056–8063, May 2009.
[37] K. Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification," Engineering Applications of Artificial Intelligence, vol. 24, no. 5, pp. 796–802, Aug. 2011.
[38] S. Russell and P. Norvig, Artificial Intelligence: Pearson New International Edition. Pearson Higher Ed, 2013.
[39] J. Kennedy, J. F. Kennedy, R. C. Eberhart, and Y. Shi, Swarm Intelligence. Morgan Kaufmann, 2001.