Parkinson's Disease Classification using Gaussian Mixture Models with Relevance Feature Weights on Vocal Feature Sets

Ouiem Bchir
College of Computer and Information Sciences
Computer Science Department, King Saud University
Riyadh, Saudi Arabia

***Abstract*—In order to perceive automatically the manifestation of dysarthria in Parkinson's disease, we propose a novel classifier which is able to categorize acoustic features and detects articulatory deficits. The proposed approach incorporates relevance feature weighting to the Gaussian mixture model in order to address the issue of high dimensionality. Besides, it learns the relevance feature weights with respect to each model along with the Gaussian mixture model parameters to deal with the specificity of the class models. In order to assess the performance of the proposed approach, we used the data collected by the department of neurology in Cerrahpaşa faculty of medicine at Istanbul University. The obtained results of the Gaussian mixture models with relevance feature weights algorithm are first compared to the GMM results, and to the most recent related work. The experimental results showed the effectiveness of the proposed approach with an accuracy of 0.89 and an MCC score of 0.7.**

***Keywords*—Gaussian Mixture Models; relevance feature weights; Parkinson's disease; acoustic feature sets**

I. Introduction

Patients suffering from Parkinson's Disease (PD) show a neurological disturbance because of the devolution and death of the neurons that produce dopamine in the central nervous system. There are 7 to 10 million patients suffering from this disease in the world. Succeeding to diagnose PD at an early stage would contribute enhancing their quality of life. However, this task is tedious and the patient may be diagnosed with PD years after. Meanwhile, there is no unique commonly used diagnosis for PD which make the task even more challenging for physicians that are not expert on PD symptoms. Indeed, around 20% of PD patients are estimated to be not diagnosed yet [1].

One well known symptom of PD is a movement disorder due to the deficiency in dopamine, responsible of movement coordination. However, not only the movement of the patient is affected by the disease, but also his voice and speech since speaking involves larynx, lung and mouth mussel movements. In fact, the vocal degeneration is believed to be a common syndrome of PD disease that appears at early stage [2]. However, the vocal degeneration could not be sensed at an early stage by human ability. Rather, it could be analyzed and identified by computer based signal processing systems [3]. In fact, classification approaches of the feature extracted from a recorded speech can provide computer aided diagnosis systems that can perceive the voice degradation automatically [4]. Recently, several approaches have been reported in the literature to aid-diagnosis speech impaired diseases. These approaches are based on extracting acoustic features from the recorded speech and classifying them as PD or non PD [5], [6], [7], [8], [9], [10], [11].

Although these approaches succeeded to predict PD syndrome, the acoustic feature that is able to discriminate PD patient from non PD one, is still not characterized. Meanwhile, considering all features yields the curse of dimensionality problem. Therefore, most of previous works perform an empirical exhaustive search for the best feature-classifier combination. Another way to tackle the problem is through feature selection. Several feature selection approaches have been reported in the literature [1], [12]. For example, some of these approaches are based on performing simultaneous clustering and feature selection [13], on dropping highly correlated variable and keeping only one [14], on a logistic regression model [15], or , on a two-level hierarchical Bayesian model [16], etc. However, combining the features could be more effective than selecting a subset of them. In fact, although a certain feature can be irrelevant when compared to other features, it can contribute to the prediction. Moreover, some feature can be relevant to a certain class while not being relevant to another. Therefore, it is beneficial to have relevance feature weights with respect to each class for a better discrimination ability of the classifier.

Gaussian mixture model classifier, GMM, has been proved to be effective in many applications ([17], [18], [19]). However, in high dimension, GMM maybe not that effective. In fact, for high-dimensional data, the Gaussian distribution is very dense toward the tail. It is against the intuition, since for low dimensional data, the Gaussian distribution is dense toward the mean. This issue makes the estimation of the Gaussian mixture model parameters challenging. For this reason, the EM algorithm may fail to estimate the Gaussian mixture model parameters. Moreover, the Gaussian model parameter estimation is even more challenging when the size of the data is not large enough compared to its dimensionality. In fact, the maximum likelihood estimation MLE results in a singular covariance matrix of the Gaussian for high dimensional data which leads to the failure of the GMM. In order to alleviate this issue of high dimensional data, several

feature selection approaches have been especially devised for GMM [20]. In addition to removing irrelevant features for the purpose of improving the classifier performance, feature selection also yields a feature reduction which solves the curse of dimensionality issue. However, this kind of feature selection is crisp. The feature is either considered relevant to the application; therefore kept or it is considered irrelevant and it is discarded. However, even though a feature is considered irrelevant when compared to the other features, it may contribute to the prediction. Moreover, the features could not be equally relevant. In this case, combining the features effectively is more important than selecting a subset of them.

Feature weighting, which have been introduced mostly in the context of clustering [8], allows to combine the feature by weighting each one according to its relevance to the application. This enhances the discrimination ability of the classifiers and reduces the dimensionality without discarding any features. Moreover, the feature weights can be specific to each class model. In fact, some feature can be relevant to a certain class model while not being relevant to another. Therefore, it is beneficial to have relevance feature weights with respect to each class model for a better discrimination ability of the classifier.

The high dimensionality of the acoustic feature limits the performance of Parkinson's dysarthria recognition systems. In order to alleviate this problem, we suggest aggregating the different feature sets by introducing relevance feature weighting to the Gaussian mixture model. The proposed approach learns the relevant features and the Gaussian mixture model parameters with respect to each class.

II. Background

The statistical estimation approach, Gaussian Mixture Model, GMM, [21] approximates the probability density function, PDF, of the data using a weighted sum of Gaussian functions. The mixture of Gaussians that fits best the data is determined by a set of parameters that maximize a likelihood function. In order to estimate these parameters, the EM algorithm is used. It alternatively estimates the model parameters and the points membership likelihood.

Let x_k be a real-valued vector of length d that represents the k^{th} instance of the data of size n, v_i is the mean of the Gaussian i, A_i its covariance and φ the model parameters. The GMM can be expressed as

$$g(x_k|\varphi) = \sum_{i=1}^{C} \pi_i p_i(x_k|\varphi_i) \quad (1)$$

where C is number of considered Gaussians, $p_i(x|\varphi_i)$ is the Gaussian function i, and π_i is the ratio of $p_i(x|\varphi_i)$ in the mixture. The model parameters A_i , v_i, and πi can be determined using the maximum likelihood estimation (MLE) technique. The log of the likelihood can be expressed as

$$\sum_{k=1}^{n}\sum_{i=1}^{C} \mu_{ik}\left(log(\pi_i p_i(x|\varphi_i))\right) = \sum_{k=1}^{n}\sum_{i=1}^{C} \mu_{ik}\left(-log(\pi_i) - \frac{n}{2}log(2\pi) - \frac{1}{2}log(|A_i|) - \frac{1}{2}(x_k - v_i)A_i^{-1}(x_k - v_i)^T\right) \quad (2)$$

where μ_{ik} is the probability that x_k is assigned to the Gaussian i, and is defined as

$$\mu_{ik} = \frac{\pi_i p_i(x_k|\varphi_i)}{\sum_{l=1}^{C} \pi_l p_l(x_k|\varphi_l)} \quad (3)$$

It can be proven that the MLE parameters are

$$v_i = \frac{1}{n}\sum_{k=1}^{n} \mu_{ik} x_k, \text{ and} \quad (4)$$

$$A_i = \frac{1}{n}\sum_{k=1}^{n} \mu_{ik}(x_k - v_i)(x_k - v_i)^T \quad (5)$$

As mentioned above, the resulting algorithm alternates the E-step and The M-step. The E-step assigns each point x_k to a Gaussian i, and the M-step computes the Gaussian centers, v_i, and covariance, A_i . The GMM algorithm using MLE optimization is summarized in algorithm 1.

Algorithm 1 GMM algorithm

Initialize v_i, A_i, and π_i
Repeat
1- E-step: Compute μ_{ik} using (3)
2- M-step:
- Compute v_i using (4)
- Compute A_i using (5)
- Compute $\pi_i = \frac{n_C}{n}$, where $n_C = \sum_{k=1}^{n} \mu_{ik}$

Until convergence

The GMM algorithm may be prone to local minima. That is why it is preferable to run it several times with different initialization settings.

III. Related Works

Recently, considerable researches that tackle the relation between PD and speech disorder have been reported in literature. More specifically, classification based approaches that recognize PD voice impairment symptom have been proposed. Their performance depends highly on the selection of the appropriate acoustic feature and on the machine learning approach adopted.

The authors in [5] suggested to select the 10 most uncorrelated features by applying redundant feature filter. Then, using the obtained features, they performed an exhaustive search looking for all possible combinations. These feature combinations are conveyed to a kernel SVM classifier to conclude on the best combination of features. They found out that the combination of pitch period entropy (PPE) [5]

and the harmonics-to-noise ratios gave the best performance. In the same context of feature selection, the authors in [6] use 22 acoustic features as described in [22]. Based on the obtained 132-length feature vector, they compared four feature selection algorithms. Namely, they used the least absolute shrinkage and selection operator (LASSO) [23], the minimum redundancy maximum relevance (mRMR) [24], the RELIEF [25] and the local learning-based feature selection (LLBFS) [26]. The empirical comparison concluded that RELIEF [25] is more suitable for this data when reducing the feature's dimension to 10. Then, the obtained 10 pre-selected features are conveyed to random forests (RF) and support vector machines (SVM) binary classifiers [27]. They concluded that SVM outperforms RF for this data.

Other researches tackled the problem by introducing new feature extraction approaches. The authors in [9] presented a system for PD system based on segmenting 'pa', 'ta', and 'ka' syllables. Using the obtained syllables, they designed 13 acoustic features to detect voice deficiency. The extracted features are then classified using SVM [27] in order to discriminate between PD and non PD patients. On the other hand, the authors in [28] applied a combination of Mel-frequency cepstral and of tunable Q-factor wavelet coefficient as a feature to be fed to a voice based PD diagnosis system. The obtained feature is conveyed to 9 classifiers that are combined using ensemble learning method.

Since one of the characteristics of the voice data for PD detection is the record repetition of the same patient, the authors in [10] and [11] proposed two systems to handle the data repetition problem. The first proposed approach is based on aggregating the data while the second one used latent variable in the Bayesian logistic regression approach. Similarly, the authors in [7] dealt with the problem of repeated voice recordings per patient. They suggested representing the acoustic features extracted from the records of the same patient with center and dispersion variables rather than with independent variables. They used the k-nearest neighbor (k-NN) and support vector machines (SVM) [27] as classification approaches to segregate between PD and non PD patients. Whereas, the authors in [29] don't address only the problem of within-patient variability but also multicollinearity. They proposed a two stage approach. The first step is a feature selection step. For each group of feature, one representative is kept based on its similarity with the feature of the same group. The second step consists in using Least Absolute Shrinkage and Selection Operator LASSO [30] that performs regression and variable selection. Moreover, Gibbs sampling algorithm [31] is used in order to avoid the computational complexity of the two stage system.

In the context of feature weighting, the authors in [8] proposed a hybrid system to detect PD from acoustic features. They first weighted the features by clustering the data using Gaussian mixture model GMM [21]. Then, they performed feature reduction and transformation using principal component analysis, PCA, linear discriminant analysis LDA, sequential forward selection SFS, and sequential backward selection SBS [32]. Finally, they classified the transformed acoustic features using least-square support vector machine LS-SVM [33], probabilistic neural network PNN [34] and general regression neural network GRNN [35]. Similarly, for feature selection purpose, the authors in [36] used recursive feature elimination algorithm (RFE) [37]. The obtained selected features were conveyed to a linear SVM classifier in order to distinguish PD from non PD patients.

Recently, the authors in [38] employed the deep learning framework to discriminate between PD and non PD patients. In fact, they introduced two systems based on Convolutional Neural Networks CNN [39] to combine several acoustic features. The first proposed system aggregates the considered features before conveying them to a CNN with 9 layers. Whereas, the second proposed system conveyed directly the considered feature to a CNN with parallel input layers. They concluded that the second system is promising.

In summary, recent researchers found that voice degeneration allowed the early diagnosis of PD. In this context, several systems based on feature extraction and machine learning methods have been reported in the literature. Some of these works ([5], [6]) focused in the problem of acoustic feature high correlation and suggested the use of different feature selection approaches. Other works ([9], [28]) introduced new feature extraction method to discriminate PD from non PD patient using voice records. The works ([7], [10], [11]) tackled the problem of repeated voice recordings per patient. Feature weighting has been considered in multiple layered hybrid system where the feature weighting is performed through clustering the data [8]. Deep learning framework has also been considered in [38] where CNN has been used to classify PD and non PD patients.

IV. Proposed Approach

Let x_k be a real-valued vector of length d that represents the k^{th} instance of the data of size n, x_k can be seen as a set of sub-vectors where each sub-vector represents a different feature. Let St be the number of considered sub-features, x_k can be expresses as

$$x_k = [x_k^1, x_k^2, x_k^3, \dots, x_k^s] \tag{6}$$

where x_k^s represents the sub-feature s of size z^s. It follows that the size d of x_k is

$$d = \sum_{s=1}^{St} z^s \tag{7}$$

Let w_{is} be the weight of sub-feature x_k^s with respect to Gaussian i. Concatenating the different sub-features to reconstruct the vector x_k would result in a high dimensional vector which yield all the related drawbacks. To alleviate this problem, we propose aggregating the different sub-features as a weighted sum of the different sub-features. The weights related to each sub-feature s are learned in such a way they reflect the relevance of sub-feature s in modeling x_k with Gaussian i. In this sense, the distance between x_k and the center of Gaussian i , can be defined as

$$\sum_{s=1}^{St} w_{is}^q (x_k^s - v_i^s) A_i^{s^{-1}} (x_k^s - v_i^s)^T \tag{8}$$

where w_{is} is the relevance weight of sub-feature s with respect to Gaussian model i, q is the parameter that controls the fuzziness of these feature relevance weights, A_i^s and v_i^s are respectively the covariance and mean of sub-feature s with respect to model i. The definition of the new distance in (8) yields, that the i^{th} Gaussian g_i, can be defined as

$$g_i(x_k|\varphi_i) = \frac{1}{(2\pi)^{n/2}\sum_{s=1}^{St} w_{is}^q \left|A_i^s\right|^{1/2}} \cdot e^{\left[-1/2\sum_{s=1}^{St} w_{is}^q (x_k^s - v_i^s)A_i^{s^{-1}}(x_k^s - v_i^s)^T\right]} \quad (9)$$

subject to:

$$\sum_{s=1}^{St} w_{is} = 1 \quad (10)$$

Let $\varphi = [\varphi_i]_{1..C}$ be the model parameters. The GMM with relevance feature weights can be expressed as

$$g(x_k|\varphi) = \sum_{i=1}^{C} \pi_i g_i(x_k|\varphi_i) \quad (11)$$

where C is number of considered Gaussians, $g_i(x|\varphi_i)$ is the Gaussian function i, and π_i is the ratio of $g_i(x|\varphi_i)$ in the mixture. The model parameters $\varphi = [v_i^s, A_i^s]_{\substack{i=1..C \\ s=1..St}}$, can be determined using the maximum likelihood estimation (MLE) technique. The logarithm of the likelihood L can be expressed as

$$L = \sum_{k=1}^{n}\sum_{i=1}^{C} (\log(\pi_i g_i(x|\varphi_i)))\, \mu_{ik} \quad (12)$$

where μ_{ik} is the probability that x_k is assigned to the Gaussian i, and is defined as

$$\mu_{ik} = \frac{\pi_i g_i(x_k|\varphi_i)}{\sum_{l=1}^{C} \pi_l g_l(x_k|\varphi_l)} \quad (13)$$

Substituting (9) in (12), gives

$$L = \sum_{k=1}^{n}\sum_{i=1}^{C} \mu_{ik}\left(-\log(\pi_i) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{s=1}^{St}\log(|A_i^s|) - \frac{1}{2}\sum_{s=1}^{St} w_{is}^q (x_k - v_i^s)A_i^{s^{-1}}(x_k - v_i^s)^T\right) \quad (14)$$

The derivative of L with respect to v_i is

$$\frac{\partial L}{\partial v_i^s} = -\frac{1}{2}\sum_{k=1}^{n} \mu_{ik} w_{is}^q (x_k - v_i^s)A_i^{s^{-1}} \quad (15)$$

Setting (15) to zero gives the estimated value of v_i as in (16)

$$v_i^s = \frac{1}{n}\sum_{k=1}^{n} \mu_{ik} w_{is}^q x_k^s \quad (16)$$

The partial derivative of L with respect to A_i^{-1} is

$$\frac{\partial L}{\partial A_i^{-1}} = -\frac{n}{2}A_i^s - \frac{1}{2}\sum_{k=1}^{n} \mu_{ik} w_{is}^q (x_k^s - v_i^s)(x_k^s - v_i^s)^T \quad (17)$$

Setting (17) to zero gives

$$A_i^s = \frac{1}{n}\sum_{k=1}^{n} \mu_{ik} w_{is}^q (x_k^s - v_i^s)(x_k^s - v_i^s)^T \quad (18)$$

Using the Lagrange multiplier technique, the partial derivative of L with respect to w_{is} subject to (10) is

$$\frac{\partial L}{\partial w_{is}} = -\frac{1}{2}\sum_{k=1}^{n} q\mu_{ik} w_{is}^{q-1}(x_k^s - v_i^s)A_i^{s^{-1}}(x_k^s - v_i^s)^T + \lambda \quad (19)$$

where λ is the Lagrange coefficient. Setting (19) to zero gives

$$w_{is} = \left(\frac{\lambda/q}{\frac{1}{2}\sum_{k=1}^{n} \mu_{ik}(x_k^s - v_i^s)A_i^{s^{-1}}(x_k^s - v_i^s)^T}\right)^{1/q-1} \quad (20)$$

Substituting (20) in (10), yields

$$\left(\lambda/q\right)^{1/q-1} = \frac{1}{\sum_s^{St}\left(\frac{1}{\frac{1}{2}\sum_{k=1}^{n} \mu_{ik}(x_k^s - v_i^s)A_i^{s^{-1}}(x_k^s - v_i^s)^T}\right)^{1/q-1}} \quad (21)$$

Substituting (21) in (20), gives

$$w_{is} = \left(\frac{1/\sum_{k=1}^{n} \mu_{ik}(x_k^s - v_i^s)A_i^{s^{-1}}(x_k^s - v_i^s)^T}{\sum_{l=1}^{St}\left(1/\sum_{k=1}^{n} \mu_{ik}(x_k^l - v_i^l)A_i^{l^{-1}}(x_k^l - v_i^l)^T\right)}\right)^{1/q-1} \quad (22)$$

As mentioned above, the resulting algorithm alternates the E-step and the M-step. The E-step assigns each point x_k to a Gaussian i, and the M-step computes the Gaussian centers, v_i^s, and covariance, A_i^s . The GMM with relevance feature weights algorithm using MLE optimization is summarized in algorithm 2.

Algorithm 2 GMM with relevance feature weights algorithm

Initialize v_i^s, A_i^s, and π_i and w_{is}
Repeat
1- E-step: Compute μ_{ik} using (13)
2- M-step:
- Compute v_i^s, using (16),
- A_i^s using (18)
- Compute w_{is} using (22)
- Compute $\pi_i = \frac{n_C}{n}$, where $n_C = \sum_{k=1}^{n} \mu_{ik}$

Until convergence

Similar to the GMM algorithm, the proposed GMM with relevance feature weights may be prone to local minima. One way to alleviate this problem is by running it several times with different initialization settings.

V. Experiments

In order to assess the performance of the proposed approach, we used the data set available at [40]. It was collected by the Department of Neurology in Cerrahpaşa Faculty of Medicine, Istanbul University. The data was built with the participation of 252 persons which age varies from 33 to 87. From the 252 participants, 188 are diagnosed with PD and 64 are healthy. The participants were asked to pronounce the vowel /a/ three times. From the recorded 756 sample vocal data, acoustic features are extracted. Namely, Time Frequency Features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform based Features, Vocal Fold Features and the tunable Q-factor wavelet transform (TQWT) features were extracted from the collected data [28]. This results in 752 dimensional feature vector for each record. The various extracted feature sets and their corresponding dimensions are reported in Table I.

For the purpose of assessing the performance of the proposed approach, the extracted feature subsets (refer to Table I) are conveyed to the Gaussian Mixture model with relevance feature weights (as described in section4). The obtained results are first compared to the GMM results. Then, they are compared to the most recent related work that uses the same data set with same extracted set of features. Namely, we compare the obtained results to those reported in [28]. More specifically, the obtained results are compared to the results obtained when conveying the top 50 features selected using mRMR to different classifiers and the combination of their prediction using ensemble stacking and voting approaches. For this purpose, two performance measures are computed. These are the accuracy, and the Matthews correlation coefficient (MCC). For both GMM and GMM with relevance feature weights, we use two models for the PD class and 2 models for the Non PD class. Since both classifiers are prone to local minima, the experiment is run 100 times. Moreover, we use the 10-cross validation technique. Table II shows the comparison of the performances of GMM and GMM with relevance feature weights. The reported results are the mean and standard deviation of the accuracy and the MCC score over the 100 runs. As it can be seen, GMM performs poorly on this data. This is due to the high dimensionality of the data. On the other hand, by learning relevance feature weights that allow an effective combination of the feature subsets, the proposed GMM with relevance feature weights overcomes the high dimensionality problem, and give better results. In order to further investigate the obtained results, we report in Table III the confusion matrix obtained using GMM with relevance feature weights, and in Table IV, the confusion matrix obtained using GMM. We notice that GMM classifies the whole data as Non PD. In fact, since the feature vector has high dimensionality, the covariance matrix learned by GMM would be singular or nearly singular resulting in the numerical breakdown of the model.

Fig. 1 depicts the learned relevance feature weights. Since we used two models for the PD class and two models for the non PD class, the proposed classifier learns a feature weight for each subset with respect to each of the 4 models. As it can be seen from Fig. 1, the first model of the non PD class has the large weight with respect to the Detrended fluctuation analysis feature, whereas the second model has the large weight with respect the Recurrence Period Density entropy feature. This means that the former feature allows discriminating the first model while the latter allows discriminating the second model. Similarly, the Pitch Period entropy, the Mel frequency features, and the vocal fold features are relevant to the first model of the PD class, while the Recurrence Period Density is relevant to the second one. By learning the relevant feature weight for each model, the proposed approach allows an effective combination of the feature subsets resulting in the improvement of the GMM performance.

TABLE I. Overview of the Feature Sets used and their Corresponding Dimensions

Feature subset	Size
Jitter variants	5
Shimmer variants	6
Fundamental frequency parameters	5
Harmonicity parameters	2
Recurrence Period Density entropy	1
Detrended Fluctuation analysis	1
Pitch Period entropy	1
Time frequency features (intensity, Frequencies, and bandwidth)	11
Mel Frequency Cepstral Coefficient	84
Wavelet transform	182
Vocal fold features (Glottis, Glotal to noise Excitation, vocal Fold Excitation, Empirical Mode Decomposition)	22
TWQT	432

TABLE II. Comparison of the performances of GMM and GMM with relevance feature weights

	Accuracy	MCC
GMM	0.2540 ± 0	0±0
GMM with relevance feature weights	0.8912± 0.0054	0.7060±0.0143

TABLE III. Confusion matrix obtained using GMM with relevance feature weights

	Predicted Non PD	Predicted PD
Actual Non PD	119	73
Actual PD	0	564

TABLE IV. Confusion matrix obtained using GMM

	Predicted Non PD	Predicted PD
Actual Non PD	192	0
Actual PD	564	0

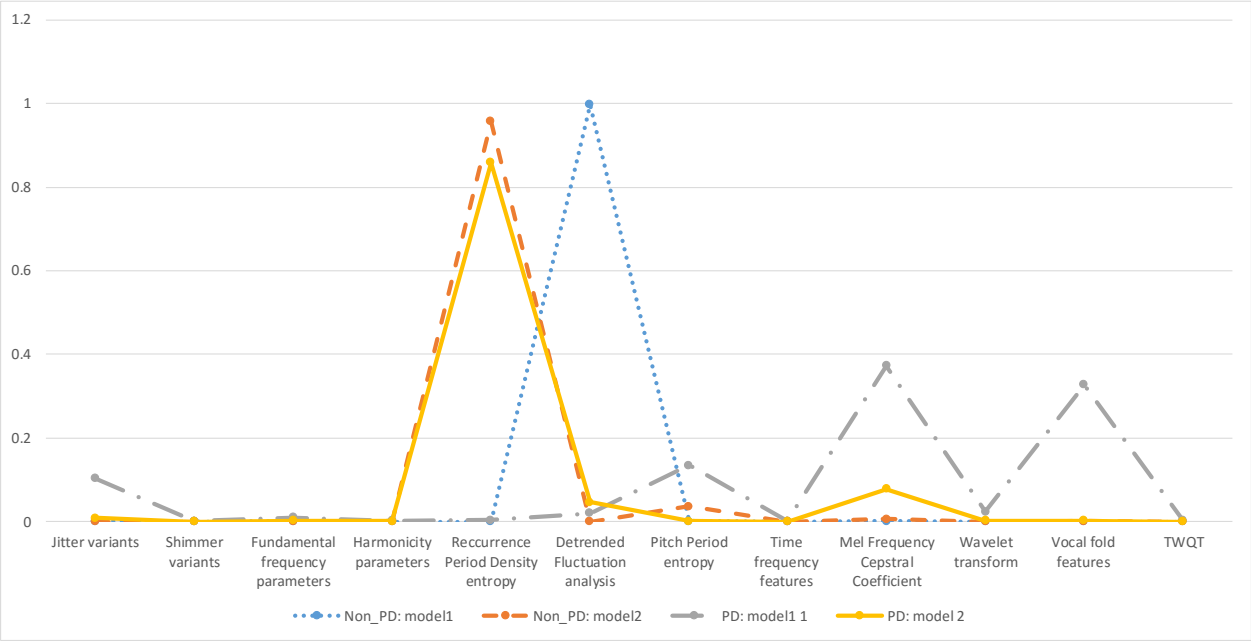

Fig 1. Relevance Feature Weights Learned by the Proposed Approach.

TABLE V. COMPARISON OF THE PERFORMANCE OF THE PROPOSED APPROACH WITH RELATED WORKS

	Accuracy	**MCC**
Naïve Bayes	0.83	0.54
Logistic regression	0.85	0.57
K-NN	0.85	0.56
Multi-layer perceptron	0.84	0.54
Random Forest	0.85	0.57
SVM (Linear)	0.83	0.52
SVM(RBF)	0.86	0.59
Ensemble with voting	0.85	0.58
Ensemble with stacking	0.84	0.55
GMM	0.2540 ± 0	0 ±0
GMM with relevance feature weights	**0.8912** ± 0.0054	**0.7060** ±0.0143

In Table V, we show the comparison of the performance of the proposed approach with the results of different classifiers on the top 50 features selected using mRMR and the combination of their prediction using ensemble stacking and voting approaches as reported in [28] on the same data set. We notice that the proposed approach has a higher accuracy and a higher MCC. This means that is outperforming the other considered approaches. We should mention here that, for the PD classification problem, higher accuracies than the accuracy of the proposed approach have been reported for the same dataset. However, these approaches use leave-one-out cross validation, whereas the dataset has several recordings per person. This yields biased models since recordings of the same person of the test recording are included in the training set, which results in overfitting problem.

VI. CONCLUSION

Recently, the number of PD patients has increased. Nowadays, 2 to 3% of older people that are over 65 years are affected by the disease. With the progression of the disease, different symptoms appear affecting the speech. Machine learning techniques can be used for early detection of PD syndromes. More specifically, speech pattern detection approaches have been applied to the problem of articulatory deficits caused by PD. Although machine learning techniques have been proven to be effective to predict PD syndrome, the relevant acoustic feature that allow to distinguish the vocal records of PD patient from non PD one, is still not solved. In fact, since considering all features would result on the curse of dimensionality problem, most of previous works compare empirically these features with different classifiers in order to come up with best combination feature-classifier. Other approaches used feature selection techniques in the whole set of features in order to reduce the dimensionality and keep only relevant features. However, even though a feature is considered irrelevant when compared to the other features, it may contribute to the prediction. In this case, combining the features effectively is more important than selecting a subset of them. Moreover, the feature selection is done for the whole data set while some feature can be relevant to a certain class while not being relevant to another. Therefore, it is beneficial to have relevance feature weights with respect to each class for a better discrimination ability of the classifier.

In this work, we classify the voice records for PD patient detection. For this purpose, we introduced a new classifier that incorporates relevance feature weighting to the Gaussian mixture models classifier. In fact, the proposed classifier learns the relevant feature weights and the Gaussian mixture model parameters with respect to each class. The experimental results showed the effectiveness of the proposed approach with an accuracy of 0.89 and an MCC score of 0.7.

As future work, we intend to combine the Gaussian Mixture model classifier with a clustering algorithm that learn the relevance feature weights.

ACKNOWLEDGMENT

The authors are grateful for the support of the Research Center of the College of Computer and Information Sciences, King Saud University. This research received no external funding.

REFERENCES

[1] M. Kyung, J. Gill, M. Ghosh, G. Casella, "Penalized regression, standard errors, and Bayesian LASSOS," Bayesian Anal. 5 (2), p. 369–412, 2010.

[2] J. R Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Elsevier, 2005.

[3] B.T. Hare, M.S. Cannizzaro, H. Cohen, N. Reilly, P.J. Snyder, "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment," J. Neurolinguistics, vol 17 (6), pp. 439-453, 2004.

[4] J. Russ, R. Cmejla, H. Ruzickova , E. Ruzicka, "Quantitative acoustic measure- ments for characterization of speech and voice disorders in early untreated Parkinson's disease," J. Acoust. Soc. Am. 129 (1), p. 350–367, 2011.

[5] M.A. Little, P.E. McSharry , E.J. Hunter , J. Spielman , L.O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Trans. Biomed. Eng. 56 (4), p. 1015–1022 ., 2009.

[6] A . Tsanas, M.A . Little , P.E. McSharry , J. Spielman , L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Trans. Biomed. Eng. 59 (5), p. 1264–1271, 2012.

[7] B.E. Sakar, M.E. Isenkul , C.O. Sakar , A. Sertbas , F. Gurgen , S. Delil , H. Apaydin , O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," IEEE J. Biomed. Health Inf. 17 (4), p. 828–83, 2013.

[8] M. Hariharan, K. Polat , R. Sindhu, "A new hybrid intelligent system for accurate detection of Parkinson's disease," Comput. Methods Programs Biomed. 113 (3), p. 904–913 ., 2014.

[9] M. Novotny, J. Rusz , R. Cmejla , E. Ruzicka, "Automatic evaluation of articulatory disorders in Parkinson's disease," IEEE/ACM Trans. Audio Speech Lang. 22 (9), p. 1366–1378 ., 2014.

[10] C.J. Pérez, L. Naranjo , J. Martín , Y. Campos-Roca, "A latent variable-based Bayesian regression to address recording replication in Parkinson's disease," in the 22nd European Signal Processing Conference, Lisbon,Portugal, 2014.

[11] L. Naranjo, C.J. Pérez , Y. Campos-Roca , J. Martín, "Addressing voice recording replications for Parkinson's disease detection," Expert Syst. Appl. 46, p. 286–292, 2016.

[12] V. Rockova, E. Lesaffre , J. Luime , B. Löwenberg, "Hierarchical Bayesian for mulations for selecting variables in regression models,," Stat. Med. 31, p. 1221–1237 ., 2012.

[13] S.M. Curtis, S.K. Ghosh, "A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression," J. Stat. Theory Pract. 5 (4), p. 715–735 ., 2011.

[14] H. Midi, S.K. Sarkar , S. Rana, "Collinearity diagnostics of binary logistic regres- sion model," J.Interdiscip. Math. 13 (3), p. 253–267., 2010.

[15] X. Zhou, K.-Y. Liu , S.T.C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," J. Biomed. Inf. 37, p. 249–259., 2004.

[16] K. Bae, B.K. Mallick, "Gene selection using a two-level hierarchical Bayesian model," Bioinformatics 20 (18), p. 3423–3430, 2004.

[17] M.S. Allili, D. Ziou, N. Bouguila, S. Boutemedjet, "Image and video segmentation by combining unsupervised generalized Gaussian mixture modeling and feature selection," IEEE Trans Circuits Syst Video Technol 20(10), p. 1373–1377, 2010.

[18] J. Tao, N. Shu, Y. Wang, Q. Hu and Y. Zhang, "A study of a Gaussian mixture model for urban land-cover mapping based on VHR remote sensing imagery,," International Journal of Remote Sensing 37(1), pp. 1-13,, 2016.

[19] I. Prabakaran, Z. Wu Z, C. Lee, B. Tong, S. Steeman, G. Koo, P.J. Zhang, M.A. Guvakova, "Gaussian Mixture Models for Probabilistic Classification of Breast Cancer 79(13)," Cancer Res., MA, pp. 3492-3502, 2019.

[20] S. Beling, and P.A Adams, "A survey of feature selection methods for Gaussian mixture models and hidden Markov models," Artificial Intelligence Review,52(3), p. 1739, 2019.

[21] C. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, New York: Springer-Verlag, 2006.

[22] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," J. Roy. Soc., vol. 8, p. 842–855, 2011.

[23] R. Tibshirani, "Regression shrinkage and selection via the Lasso," J. Roy.Statist. Soc. B, vol. 58, p. 267–288, 1996.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and minredundancy," IEEE Trans. Pattern Anal. Mach 27(8), p. 1226–1238, 2005.

[25] K. Rendell, and L.A Kira, "A practical approach to feature selection," in 9th Int. Conf. Mach. Learn., 1992.

[26] Y. Sun, S. Todorovic, and S. Goodison, "Local learning based feature selection for high dimensional data analysis," IEEE Pattern Anal. Mach., 32(9), p. 1610–1626, 2010.

[27] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, New York: Springer, 2009.

[28] C. O. Sakar, G. Serbe, A. Gunduz , H. C. Tunc , H. Nizam , B. Sakar , M. Tutuncu , T. Aydin , M. E. Isenkul , H. Apaydin, "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the use of the tunable Q-factor wavelet transform," Applied Soft Computing,74, pp. 255-263, 2019.

[29] L. Naranjo, C. J. Pérez , J. Martín, Y. Campos-Roca, ""A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," Computer Methods and Programs in Biomedcinne 142, pp. 147-156, 2017.

[30] F. Santosa, and W. Symes, "Linear inversion of band-limited reflection seismograms," SIAM Journal on Scientific and Statistical Computing 7 (4), p. 1307–1330, 1986.

[31] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence. 6 (6), p. 721–741, 1984.

[32] V. Anuradha, and J. Bachu, "A Review of Feature Selection and Its Methods," Cybernetics and Information Technologies, volume 19, 2016.

[33] J.A.K Suykens, J. Vandewalle, "Least squares support vector machine classifiers," Neural Processing Letters, 9 (3), p. 293–300., 1999.

[34] D. F. Specht, "Probabilistic neural networks," Neural Networks (3), p. 109–118, 1990.

[35] D. F. Specht, "A general regression neural network," IEEE Transactions on Neural Networks 2(6), p. 568–576, 2002.

[36] S. Aich, M. Sain, J. Park, K. Choi and H. Kim, "A Mixed Classification Approach for the Prediction of Parkinson's disease using Nonlinear Feature Selection Technique based on the Voice Recording," in Int. Conf. on Inventive Computing and Infromatics, 2017.

[37] P.M.Granitto, C.Furlanello, F.Biasioli, and F.Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agro industrial products," Chemometrics and Intelligent Laboratory Systems, 83(2), pp. 83-90, 2006.

[38] H. Gunduz, "Deep learning-based parkinson's disease Classification using vocal feature sets," IEEE access, , special section on deep learning for computer-aided medical diagnosis, volume 7, 2019.

[39] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

[40] UCI,"https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification#," in UCI machine learning reposotory, [last visted 1/9/2020].