Multimedia Tools and Applications (2020) 79:1261–1289
https://doi.org/10.1007/s11042-019-08222-8

Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm

Fatemeh Daneshfar[1] · Seyed Jahanshah Kabudian[1]

Received: 18 April 2018 / Revised: 21 July 2019 / Accepted: 13 September 2019 /
Published online: 26 October 2019

Abstract

In recent years, Speech Emotion Recognition (SER) has received considerable attention in affective computing field. In this paper, an improved system for SER is proposed. In the feature extraction step, a hybrid high-dimensional rich feature vector is extracted from both speech signal and glottal-waveform signal using techniques such as MFCC, PLPC, and MVDR. The prosodic features derived from fundamental frequency (f0) contour are also added to this feature vector. The proposed system is based on a holistic approach that employs a modified quantum-behaved particle swarm optimization (QPSO) algorithm (called pQPSO) to estimate both the optimal projection matrix for feature-vector dimension reduction and Gaussian Mixture Model (GMM) classifier parameters. Since the problem parameters are in a limited range and the standard QPSO algorithm performs a search in an infinite range, in this paper, the QPSO is modified in such a way that it uses a truncated probability distribution and makes the search more efficient. The system works in real-time and is evaluated on three standard emotional speech databases Berlin database of emotional speech (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) and Interactive Emotional Dyadic Motion Capture (IEMOCAP). The proposed method improves the accuracy of the SER system compared to classical methods such as FA, PCA, PPCA, LDA, standard QPSO, wQPSO, and deep neural network, and also outperforms many state-of-the-art recent approaches that use the same datasets.

Keywords Speech emotion recognition · Dimension reduction · Quantum-behaved particle swarm optimization

✉ Seyed Jahanshah Kabudian
Kabudian@razi.ac.ir

[1] Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran

1 Introduction

The speech signal is the most important and natural way of communication between humans. In this communication, the speaker's emotion plays a determinative role in the transfer of concepts so that a change in the emotion can lead to different interpretations of speech. Hence, to create a perfect interaction between man and machine, the speech emotion recognition (SER) has become one of the attractive subjects for researchers. In any accurate SER system, along with the selection of important features, an efficient way to reduce the dimension of the data is required. Joint dimensionality reduction and classifier parameter estimation in SER systems can be considered as a multi-objective problem, but to the best of our knowledge, this issue has not been addressed in the literature. In this paper, a new method is proposed to detect speech emotion, using a modified QPSO algorithm for joint dimensionality reduction-classifier parameter estimation.

At the beginning step of the proposed method, Mel-Frequency Cepstral Coefficient (MFCC), Perceptual Linear Predictive Cepstral Coefficient (PLPC) and Perceptual Minimum Variance Distortionless Response (PMVDR), pitch information and their first and second-order derivatives are extracted from both speech signal and its glottal waveforms as features. Then, the usual dimensionality reduction algorithms such as Principal Component Analysis (PCA), Probabilistic Principal Component Analysis (PPCA) and Factor Analysis (FA) are applied to the extracted feature vectors and form three matrices. These three separated matrices will be used as the three particles of the initial population of the modified QPSO algorithm.

The modified QPSO algorithm is used to optimize both the projection matrix and the GMM classifier parameters. After the dimension reduction step, the Gaussian Mixture Model (GMM) is eventually trained for classifying the emotions. Due to the high correlation between glottal features, glottal waveform and emotions, and the effect of each person's emotion and his speech style on the glottal waveform [45], the glottal waveform-based features have been used in this study.

Most of the parameter estimation algorithms that are used to estimate transformation matrix may be caught up in local solutions, but the proposed method is a metaheuristic/global optimization one and therefore, is less probable to get trapped in local solutions. The objective function of this algorithm is directly the accuracy of the emotion classification on the development data, which is more effective to find dimensionality reduction parameters compared to methods with the indirect objective functions. Another advantage of the proposed algorithm is that, in the standard QPSO, each particle may be generated outside the desired range, which produces invalid solutions and the elimination or repairing that invalid solutions leads to a reduction of the convergence rate of the algorithm. To deal with this problem, in the proposed method, the truncated probability distribution is used to generate new particles in the desired range.

Details of the algorithm implementation and experimental results have been presented in the following sections. Section 2 explains the literature review. Section 3 describes the SER systems structure and the modified-QPSO method in detail. Section 4 illustrates the proposed methodology. In Section 5, the experimental setup required for the proposed strategy has been introduced, the results of comparative experiments will be presented in Section 6, and finally, the interpretation of the results will be explained to the reader in Section 7.

2 Literature review

There are many speech emotion recognition researches in recent years that have been done on emotional feature extraction [8, 24, 34, 35, 42, 63, 69, 78, 83], emotional feature

dimension reduction [14, 16, 17, 25, 26, 42, 46, 54, 59, 83, 84] and emotional feature classification [6, 8, 9, 17, 25–27, 42, 51]. Moreover, many global optimization solutions have been proposed for emotion recognition [1, 10, 18, 22, 44, 70, 76] until now. Darekar and Dhande [10] has proposed an adaptive learning architecture for the artificial neural network to learn the multimodal fusion of speech features using a hybrid particle swarm optimization (PSO) algorithm. A facial expression recognition system using evolutionary particle swarm optimization-based feature optimization has been introduced by Mistry et al. [44]. Similarly, in [70], Wang et al. proposed a novel intelligent emotion recognition system that used stationary wavelet entropy to extract features, and employed a single hidden layer feedforward neural network as the classifier. Likewise, Albornoz et al. [1] used an auditory signal representation to obtain a novel bio-inspired set of features for emotional speech signals. Moreover, Gharavian et al. [18] have employed the particle swarm optimization to determine the optimum values of chosen parameters of extracted features. In [22] Yogesh et al. proposed a new particle swarm optimization-assisted biogeography-based algorithm for feature selection, and finally Yogesh et al. [76] have employed a biogeography-based optimization, particle swarm optimization and a proposed BBO_PSO hybrid optimization for feature selection.

The summary of other currently published works on SER systems is illustrated in Table 1.

3 Background

3.1 Speech emotion recognition systems

SER systems indicate systems with a speech signal as input and estimated emotion as the output. Like many pattern recognition systems, these systems approximately characterize the emotion of the input signal, based on the signal features and classification.

Generally, a typical SER system consists of four different parts: preprocessing, feature extraction, dimension reduction (optional) and feature classification (Fig. 1). At first, the preprocessing of the speech signal before feature extraction has been considered. It is an important stage of an efficient speech emotion recognition system. Pre-emphasizing, framing, windowing, and voice activity detection are three common techniques used in signal preprocessing. The preprocessed signal will be then fed to the feature extraction module. In this stage, the necessary and emotion-relevant features will be extracted from the signal. These features can be categorized in three different kinds [8], prosodic features, such as pitch and energy, spectral features, such as formants, MFCC, and linear predictive cepstral coefficients (LPCC), and voice quality features. Then a feature vector corresponding to each frame will be prepared. For reducing the redundancy and dimensionality of the generated feature vector, feature dimensionality reduction is needed. The dimension reduction module gets the redundant high-dimensional feature vector as input and reduces the feature vector dimensionality by applying and affixing transformation (projection) matrix. There are so many feature dimensionality reduction solutions including unsupervised dimension reduction methods (such as FA, PPCA, ICA, CCA) and supervised ones (such as Linear Discriminant Analysis (LDA)). Finally, in the last part of the system, a classification method like a neural network, support vector machine, Gaussian mixture model, hidden Markov model, etc. will classify the dimensionality-reduced features and estimate the emotional class of the input signal.

Table 1 Currently state of the art works on speech emotion recognition

SER Stages	Method	Year	Reference
Emotional feature extraction	Utilized inherent long-term properties of acoustic features by a modulation filtering approach	2014	[52]
	Captured the deviations in features related to the excitation source component of speech	2015	[30]
	Proposed a discriminant analysis based on a deep neural network to learn discriminative features	2011	[61]
	Proposed the pH time-frequency vocal source feature	2014	[79]
	Incorporated rhythm and temporal features	2012	[4]
	Ranked and selected features by their Fisher discriminant ratios	2016	[41]
	Using prosody, spectral envelope, and voice quality features	2010	[40]
	Using a deep convolutional neural network to extract spectrograms features	2019	[2]
	Using both prosodic and spectral features by Naïve Bayes Classifier	2017	[33]
	Using biologically-inspired auditory attention features	2016	[31]
	Proposed a new feature, residual sinusoidal peak amplitude	2016	[11]
	Employed the power-normalized cepstral coefficients as features	2016	[3]
	Using the phase of the pitch harmonic as feature	2017	[12]
	Provided deep belief network to yield the higher-level features from the low-level features	2017	[71]
	Proposed new features based on the energy content of wavelet-based time-frequency analysis	2016	[67]
	Provided statistics of pitch and energy as well as spectral features	2015	[58]
	Extracted spectrogram features from the speech and glottal flow signals	2016	[19]
	Applying a deep convolutional neural network to speech spectrograms	2018	[37]
	Proposed the learned deep spectrum features	2018	[82]
	Utilized spectral, prosody and voice quality features	2015	[36]
	Using low-level descriptors (local features) and statistical functional (global features)	2017	[65]
Emotional feature dimension reduction	Proposed a multiscale kernel for feature reduction	2015	[73]
	Applied a nonlinear dimensionality reduction method	2013	[80]
	Applied nuisance attribute projection to project the emotion vectors to a minimum subspace	2016	[41]
	Utilized support vector machine	2016	[56]
	Presents a semi-supervised feature selection method	2015	[62]
Emotional feature classification	Utilized SVM for classification	2017 2016 2014	[74] [56] [77]
	Using a binary decision tree for classification	2014	[77]
	Using two linear and Gaussian radial basis function kernels with binary tree	2015	[58]
	Proposed an ensemble softmax regression model	2017	[63]
	Using a deep neural network	2017	[53]

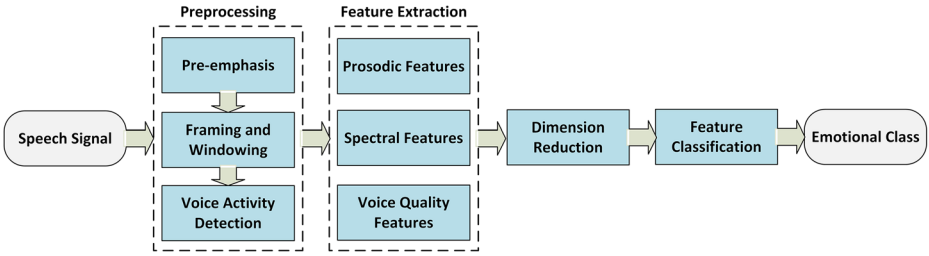


Fig. 1 The hierarchy of a Speech Emotion Recognition (SER) system

3.2 The proposed point mass function-weighted QPSO (pQPSO)

Optimization algorithms include all algorithms which try to find a locally- or a globally-best solution that optimizes a predefined objective function (Fig. 2). In this category, metaheuristic algorithms including nature-inspired swarm-based optimization, have become increasingly popular. On the other hand, quantum mechanics-based concepts have been applied to many metaheuristic optimization algorithms like genetic algorithm and particle swarm optimization [64]. For example, quantum-behaved particle swarm optimization has fewer parameters to be adjusted compared to its classic version.

In the classical PSO, each particle has a position and a velocity [64]. However, in the QPSO algorithm according to the uncertainty principle, the particle position and velocity cannot be determined simultaneously [49]. Thus it will be computed by a probability density function, that determines the probability of a particle appearing in a position and in a time [72],

Point Mass Function-weighted QPSO (called pQPSO) is a modified version of QPSO which has been proposed by the authors in this paper. In this modified QPSO, mean-best and global-best positions (particles) in the standard QPSO [72] are replaced by a particle which is generated based on the concept of the PMF selection (Algorithm 1) according to their relative competence in the cost function from a set of top K best particles found.

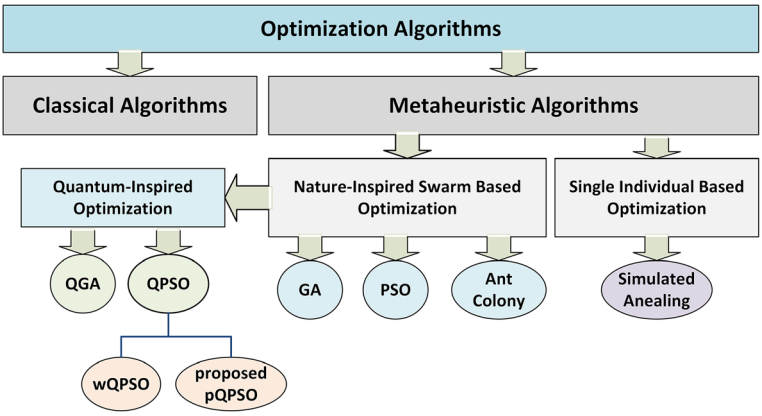


Fig. 2 The hierarchy of metaheuristic optimization algorithms

With this PMF selection, there is a chance that any of the top K best particles, which have better features or characteristics, may be selected as mean best or global best solution which reduces the greediness of the QPSO algorithm.

In the standard QPSO algorithm, new particles may be created in an invalid range and represent invalid solutions; therefore, in the proposed QPSO algorithm, in each iteration, the truncated Laplace distribution (TLD) is used to ensure that the particle values are within the corresponding valid range. Moreover, an adaptive algorithm is proposed for calculating the contraction-expansion coefficient which controls the algorithm convergence speed in each iteration, whose value is calculated proportionally to the error reduction rate in the previous iteration.

For more details on the modified QPSO algorithm, the reader can refer to [47].

Algorithm 1 Generating random m using PMF $p(m) = \sum_{j=1}^{K} w_j \delta(m - B_j)$

```
Input: {w_i}, {B_i}
Output: m
    F_0 = 0
    for i = 1 : K do
            F_i = F_{i-1} + w_i
    end for
    Generate a uniform random number u ~ U(0,1)
    for i = 1 : K do
        if F_{i-1} ≤ u ≤ F_i then
            m = B_i
        end if
    end for
```

In the following sections, the whole proposed methodology and details of SER will be explained.

4 Proposed Methodology

The whole process of the proposed methodology has been illustrated in Fig. 3. At first, silence intervals have been detected and removed in the preprocessing stage. Then the feature vector has been extracted from the input speech signal. Then the optimal dimension reduction matrix (projection matrix/affine transformation) parameters has been estimated using the pQPSO algorithm, and feature dimension reduction using the obtained optimal dimension reduction matrix has been done on both training and test feature vectors. Finally, the GMM model has been trained on the reduced-dimension train feature vectors and estimates the class of the test speech signal.

It is worth noting that all the SER systems need an emotional database for evaluating their performance (more details are in Section 5). The whole emotional database should be partitioned to the training, development and testing sets.

As it was mentioned previously, the first important part of an SER system is feature extraction. In this study since the silence intervals of the input signals do not have important effects on the expressed emotion, before the feature extraction stage; they have been detected and removed using the voice activity detection program of COVAREP toolbox [13]. After the preprocessing stage, the input wave signals are ready for feature extraction. The other steps of the solution will be described in the following sections and to be illustrated in details at Fig. 4 and Algorithm (2),

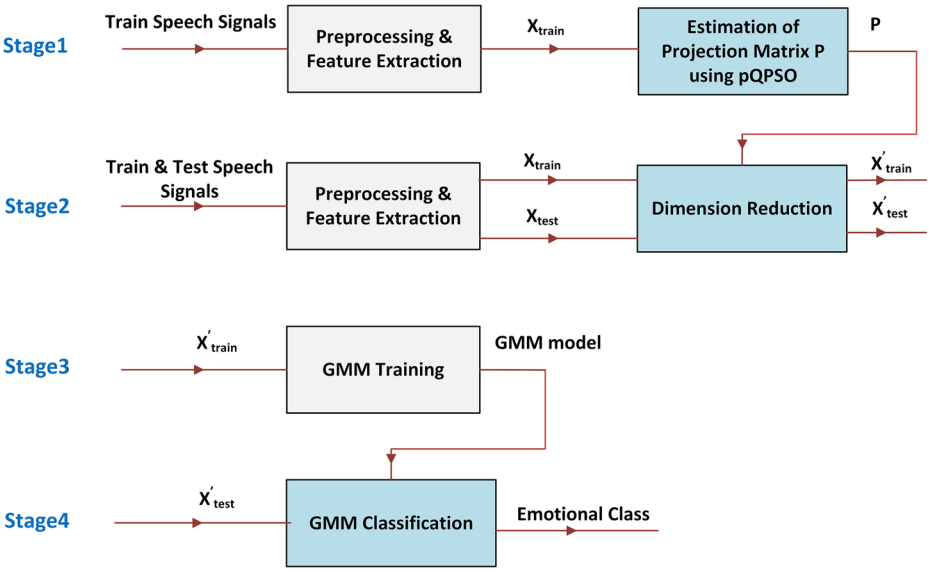


Fig. 3 The whole process flow of the proposed methodology

Algorithm 2 Whole flowchart of the proposed algorithm (equivalent to Fig. 4)

Input: training ($trnSpch$), development ($devSpch$), test ($tstSpch$) speech signals and their glottal waveforms
Output: FinalPerformance
$trnFtr = \{\}, devFtr = \{\}, tstFtr = \{\}$
% Creating train, dev and text feature vectors
for $i = 1: NumFiles$ **do**
  $trnFtr_i \leftarrow$ extract features $\{MFCC, PLPC, PMVDR, pitch + \Delta + \Delta^2\}$ from train file $trnSpch_i$
  $devFtr_i \leftarrow$ extract features $\{MFCC, PLPC, PMVDR, pitch + \Delta + \Delta^2\}$ from dev. file $devSpch_i$
  $tstFtr_i \leftarrow$ extract features $\{MFCC, PLPC, PMVDR, pitch + \Delta + \Delta^2\}$ from test file $tstSpch_i$
  $trnFtr = trnFtr \cup trnFtr_i$
  $devFtr = devFtr \cup devFtr_i$
  $tstFtr = tstFtr \cup tstFtr_i$
end for
% Learning Optimal Dimension Reduction matrix using Modified QPSO according to Algorithm3.
$OptimalDimRedMtr \leftarrow$ **Modified-QPSO-based dimension reduction** $(trnFtr, devFtr)$ *% **Algorithm3***
% Reduce dimension of training and development features
$redFtr \leftarrow OptimalDimRedMtr \times (\{trnFtr \cup devFtr\} - \mu)$
$gmm \leftarrow$ train a final GMM model $(redFtr)$
$tstRedFtr \leftarrow OptimalDimRedMtr \times (tstFtr - \mu)$
FinalPerformance = GMM-based-Classification(gmm, $tstRedFtr$)

4.1 Emotional speech feature extraction

In the context of SER systems, it is not clear yet which features most efficiently characterize various speech emotions. However, commonly frame-by-frame or short-term features extracted in the speech emotion recognition literatures include MFCC, Linear Predictive Cepstral Coefficients (LPCC), PLPC, etc. In this work, the Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Cepstral Coefficients (PLPC), Perceptual Minimum-Variance Distortionless Response Cepstral Coefficients (PMVDR), pitch (F_0) and their first- and second-order derivatives have been extracted as a feature vector for the input speech signal and also for its glottal waveform signal frame by frame (Fig. 4). Since these features are derived by different methods, they can

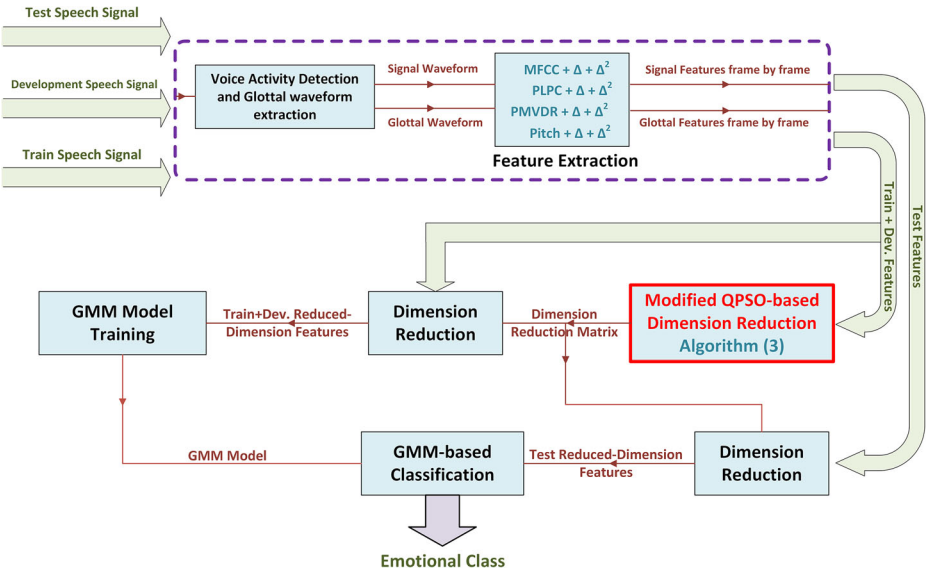


Fig. 4 Detail of the process flow of the proposed methodology (corresponding to Algorithm (2))

describe a speech signal from different aspects. All of these features were concatenated together to build a whole feature vector. In this work, MFCC and PLPC features have been extracted by Dan Ellis's toolbox [23], PMVDR was extracted using the algorithm in [75] and the pitch features were extracted using COVAREP toolbox [13]. All feature extraction levels have been carried out particularly on both input speech signal and its glottal waveform. To extract the glottal waveform signal from the speech waveform, the COVAREP toolbox [13] has been utilized again. Interestingly, experiments show that when there is a mismatch between training and testing data, the first and second-order derivatives help cope with this mismatch and improve performance in noisy environments [28, 68] so that the first and second-order derivatives can also be added to the feature vector. After all the features have been extracted from both speech and glottal waveform, a feature matrix, whose rows are frames and whose columns are of different feature vector elements, will be obtained for the next stages.

4.2 Dimension reduction

Dimension reduction aim is to reduce feature-space dimensionality to select more informative features and to reduce redundancy. Since more overlap among the features of various classes causes more performance degradation of speech emotion recognition system, dimension reduction refers to all strategies and solutions detecting a linear or nonlinear mapping between the original feature space and reduced-dimensionality space while decreasing intra-class variance and increasing inter-class distance. Therefore, if x is the input to the dimension reduction module, the output is as follows,

$$x' = P(x-\mu) \tag{1}$$

P is the dimension reduction matrix (projection matrix), and μ is the feature mean vector. The main aim of the modified QPSO-based dimension reduction box of Fig. 4, is to

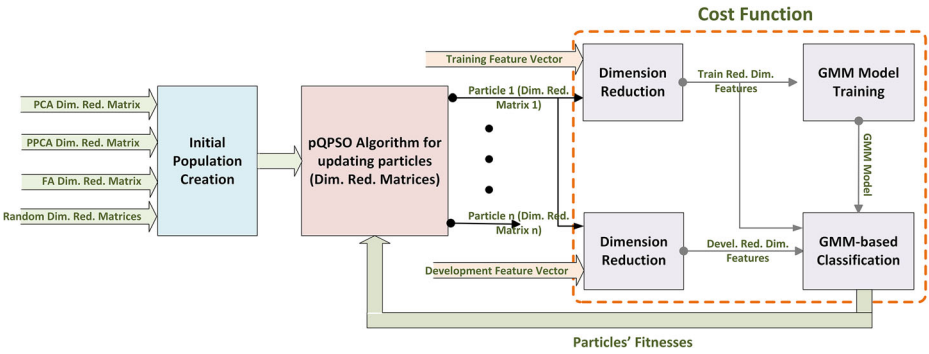


Fig. 5 The whole process flow of the pQPSO algorithm for learning an optimal dimension reduction matrix (corresponding to Algorithm (3))

optimize the dimension reduction matrix P and to find a good dimensionality-reduced feature set for classification (More information on different dimensionality reduction techniques are available in [15]).

4.2.1 Classical methods

There are many classical dimension reduction methods. Principal component analysis (PCA) is the most famous linear method for dimensionality reduction, creating a linear mapping between original features and lower-dimension features so that the variance of the low-dimensional features can be maximized [15]. If any feature vector has one or more missing values, the probabilistic principal component analysis (PPCA) is used to estimate the linear mapping. On the other hand, the factor analysis (FA) is another dimension reduction method based on the explorative analysis. It groups similar variables without distinguishing between independent and dependent variables [8].

In this research, dimension reduction matrices determined by these three famous dimension reduction solutions (i.e. PCA, PPCA and FA) will be members of the initial population of the modified-QPSO algorithm explained below. Another dimension reduction technique trying to find a discriminative dimension reduction matrix is linear discriminant analysis (LDA). However, one of the limitations of classical LDA is that the dimensionality of reduced-dimension features cannot be greater than the number of classes minus one.

4.2.2 Modified-QPSO-based strategy for dimension reduction

In this paper, a new strategy based on the QPSO algorithm has been proposed to find an optimal dimension reduction matrix. The flowchart of the proposed modified-QPSO-based dimension reduction method is illustrated in Fig. 5, and its learning scheme is summarized in Algorithm (2) and Algorithm (3). In this section, the newly-proposed pQPSO algorithm has been employed; however, other revisions of the QPSO algorithm have been employed as well and their experimental results for comparisons are presented in Section 5.

Algorithm 3 Modified-QPSO-Based Strategy for learning optimal dimension reduction matrix (Equivalent to Fig. 5)

Input: $trnFtr$, $devFtr$, and classical dimension reduction matrices ($pcaMat, ppcaMat, faMat$)
Output: Optimal Dimension Reduction Matrix ($OptimalDimRedMtr$)
% Create an initial population of particles
P_1=*Dimension Reduction Matrix found by PCA Method (Initialize First Particle)*
P_2=*Dimension Reduction Matrix found by PPCA Method (Initialize Second Particle)*
P_3=*Dimension Reduction Matrix found by FA Method (Initialize Third Particle)*
Initialize $P_4 \ldots P_{NoPopul}$ *Particles Randomly (Initialize All Other Particles (Matrices) Randomly)*
for $n = 1 : iterationNo$ **do**
    % Calculate particles' fitnesses
    for $i = 1 : NoPopul$ **do**
        $P_i' \leftarrow Householder(P_i)$
        $trnRedFtr \leftarrow P_i' \times (trnFtr - \mu)$
        $devRedFtr \leftarrow P_i' \times (devFtr - \mu)$
        $gmm \leftarrow$ train a temporary GMM model ($trnRedFtr$)
        $Fitn_i \leftarrow$ ClassificationAccuracy(gmm, $\{trnRedFtr \cup devRedFtr\}$)
    end for
    Update 'Personal Best' and 'Global Best' Positions, Update Particles Positions Using pQPSO equations
end for
% Return the 'global best' particle as 'Optimal Dimension Reduction Matrix'
$OptimalDimRedMtr = Householder$(Global Best particle of the Population)

Conventionally, there are training and development sets for training the pQPSO algorithm and for cross-validating its functionality, respectively. The pQPSO is trained on the training set, a part of the speech emotion database, and is cross-validated on both training and development sets. Finally, the evaluation is performed by a cost function.

As it is explained previously, the pQPSO algorithm aims to find an optimal dimension reduction matrix. The value of each pQPSO particle is a matrix that will be the optimal dimension reduction matrix after being optimized and after the last iteration of the pQPSO algorithm. The initial values for three pQPSO particles are dimension reduction matrices obtained by PCA, PPCA and FA methods. The other particles will be initialized randomly. After initialization, the Householder transformation [60] will be applied to each particle (matrix) to be orthogonalized and the cost function for each particle will be computed as follows.

The cost function module for each input particle finds the reduced-dimension version of both training and development feature vectors by applying that dimension reduction matrix (particle) to the feature vectors (Fig. 5 and Algorithm 3). Then, a GMM model with 128 components on the reduced-dimension training features is trained and the trained GMM is used for classifying both training and development reduced-dimension features. The correct classification rate is the fitness value for that particle. The other stages of the pQPSO algorithm were explained previously in Section 2.

4.3 Feature classification

After the pQPSO optimization has finished, it returns the global best particle as the optimal dimension reduction matrix (Fig. 4). The optimal dimension reduction matrix will be applied to training, development and test feature vector sets to reduce their dimensions. Finally, a GMM model with 128 components will be trained on both training and development reduced-dimension feature sets and will be utilized to classify it. Eventually, the final recognition rate on the test set illustrates the performance of the proposed algorithm.

5 Experimental setup

In these experiments, two different database-division strategies will be used. One of them is used for speaker-independent experiments and another one will be used for speaker-dependent experiments. Also, two different folding strategies for experiment results have been applied. In the first folding strategy (three-folds folding), the database was split into three various folds in which the training, development and testing data sets were completely distinct. Table 2 illustrates how the EMO-DB database is divided into train, development and test sets in both speaker-independent and speaker-dependent cases. In each fold, the test set contains both male and female speakers. However, in the second folding strategy (LOSO-folding), test-runs are carried out in Leave-One-Speaker-Out (LOSO) cross-validation manner just to deal with the speaker-independent case, as required by most applications. In the first folding strategy (three-folds folding), the test set includes both genders, but in the second one (LOSO folding) the test set includes only one speaker (one gender).

To validate the performance of the proposed strategy, some experiments have been setup on the following databases.

5.1 EMO-DB

The Berlin Database of Emotional Speech (EMO-DB) [5] is a German database that consists of 10 actors (five men and five women) who speak 49, 58, 43, 38, 55, 61, 69, 56 and 71 utterances, respectively. In the EMO-DB, assigned indices to these 10 actors (speakers) are 3, 8, 9, 10, 11, 12, 13, 14, 15, 16. It is one of the most popular databases used in SER systems. Each actor produces 10 routine German sentences (five short and five long utterances) in seven different emotions. The EMO-DB contains 535 distinct sentences including anger (127), fear (69), boredom (81), disgust (46), joy (71), neutral (79) and sadness (62) sentences.

5.2 SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) Database [21] consists of emotional speech British English voices from 4 male actors. It includes 480 utterances (120 utterances per actor) of 7 different emotions i.e., anger (a), disgust (d), fear (f), happiness (h), sadness (sa), surprise (su) and neutral (n).

5.3 IEMOCAP

Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [7] which was collected in 5 sessions, contains 12 h of video each of which has one female and one male speaker in both scripted and improvised scenarios. The audio files consist of 10,039 utterances produced by the English native speakers. There are nine different emotions; however, for our experiments, only improvised utterances with majority agreement including only four emotions, angry, happy, neutral and sad, where we merge excitement and happiness emotions were considered. This experimental condition was considered by many researchers [19, 36, 37, 53, 56, 65].

Table 2 Three-fold folding strategy: EMO-DB splitting into train, development and test sets in each fold, in speaker-independent and speaker-dependent experiments

	Fold 1		Fold 2		Fold 3	
	Test Data	Train and Development Data	Test Data	Train and Development Data	Test Data	Train and Development Data
Speaker-Independent	All utterances of speakers **3,8,9**	All utterances of speakers **10,11,12,13,14,15,16** (80% for Train, 20% for Dev.)	All utterances of speakers **10,11,13**	All utterances of speakers **3,8,9,12,14,15,16** (80% for Train, 20% for Dev.)	All utterances of speakers **12,14,15,16**	All utterances of speakers **3,8,9,10,11,13** (80% for Train, 20% for Dev.)
Speaker-Dependent	The **first** 30% of all utterances of every speaker	The remaining 70% of all utterances of every speaker	The **second** 30% of all utterances of every speaker	The remaining 70% of all utterances of every speaker	The **third** 40% of all utterances of every speaker	The remaining 60% of all utterances of every speaker

5.4 Number of cepstral coefficients

One of the important factors influencing system performance is the suitable number of cepstral coefficients (in MFCC, PLPC, and PMVDR features). As it was mentioned previously, there is no any comprehensive research on the best cepstral coefficients number yet [27]. However, in this study, an approximately suitable number of cepstral coefficient is proposed. The results are illustrated in Table 3.

In addition to the cepstral coefficients number, classical dimension reduction methods, i.e. PCA, PPCA, and FA, can have a different number of principal components/factors. For determining both cepstral numbers and number of principal components/factors, some experiments have been carried out by conventional dimension reduction methods such as PCA, PPCA, FA, and LDA in the same conditions in both folding strategies mentioned previously. The speech emotion recognition rate for each dimension reduction method is illustrated in Fig. 6 (a to d), and the average values have been presented in Fig. 6 (e). As is shown, the high number of cepstral coefficients and principal components does not improve the performance. Here, a relative good number of cepstral coefficients and relative good number of reduced features are equal to 7 and 20, respectively, which are achieving 69.11% for average recognition rate of PCA, PPCA and FA solutions in the three-fold folding case and 77.08% for average recognition rate of PCA, PPCA and FA solutions in the LOSO folding case of EMO-DB.

5.5 pQPSO initial population

As it was mentioned previously, the pQPSO algorithm aims to find an optimal dimension reduction matrix to reduce the dimensionality of the feature vectors. In this experiment, the initial population of the pQPSO has 40 different particles or dimension reduction matrices. In each iteration, these particles will be improved and better-fitted on the development set. The values of three members of the initial population are three dimension reduction matrices derived from the classical solutions PCA, PPCA, and FA, and the values of remaining members will be generated randomly. The PCA, PPCA and FA dimension reduction matrices in the initial population have been computed considering the previously-optimized number of cepstral coefficients and reduced-dimensionality features.

6 Results and analysis

Here the results of the proposed method in various experiments with the best-selected conditions explained in Section 5, will be explained.

Table 3 Feature extraction related parameters used in the experiments. First row: The number of cepstral coefficients. Second row: The dimensionality of total feature vector (after adding first- and second-order and concatenating all feature types (Fig. 4). Third row: The dimensionality of the final feature vector after dimension reduction using PCA, PPCA, and FA

Number of cepstral coefficients (in MFCC, PLPC & PMVDR)	7	7	7	7	13	20
The dimensionality of the total feature vector before dimension reduction	213	213	213	213	393	603
The dimensionality of the total feature vector after dimension reduction	9	20	30	40	18	24

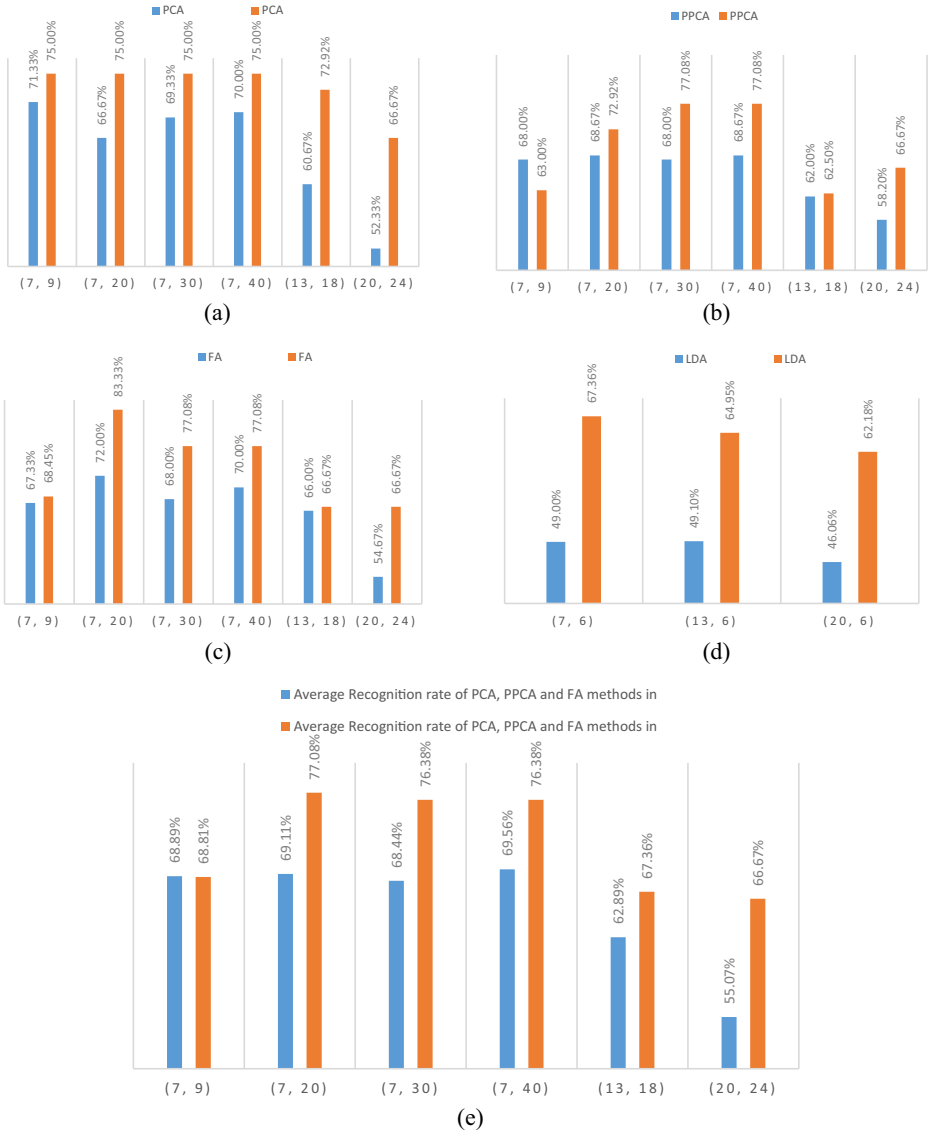


Fig. 6 The figure indicates the relation between speech emotion recognition rate and (d_1, d_2) pair (d_1 = number of cepstral coefficients, d_2 = feature dimensionality after dimension reduction) for different conventional dimension reduction methods (*a* to *d*) and their recognition rate averages (*e*). The best performance is achieved by (7,20)

6.1 First classification results

In this section, the speaker-independent (SI) and speaker-dependent (SD) scenarios were selected for experiments. At first, the classical dimension reduction algorithms such as PCA, PPCA, FA, and LDA were executed separately on the train+development sets and each dimension reduction matrix was computed. These dimension reduction matrices have been applied to the test set to compute the reduced-dimension features and measure the emotion

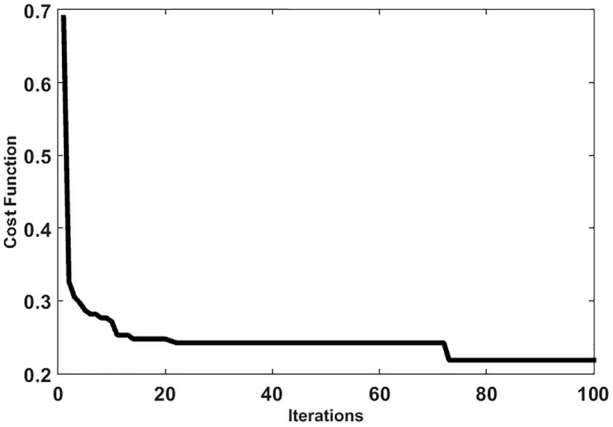


Fig. 7 The best-particle cost function value in terms of recognition error rate during pQPSO iterations (generations) on EMO-DB

recognition rate. After that, the computed dimension reduction matrices (from PCA, PPCA, and FA) were used as three different initial particles of the pQPSO algorithm and the pQPSO was run on the train and development sets as explained previously. Figure 7 illustrates the best-particle cost function values during training for 100 consecutive iterations (generations). Then, the best particle whose value or position is the best dimension reduction matrix has been applied to the test set for computing reduced-dimension features and measuring GMM-based correct classification rate.

Also, for comparing the results, the same experiments have been carried out on a deep neural network (DNN) with five hidden layers and [100,150,100 80,100] neurons in each hidden layer trained using DeeBNet toolbox [32] and on recent works [52, 74]. The average recognition rates for the three-folds folding case are illustrated in Table 4. As it is shown in both SD and SI cases, the pQPSO algorithm significantly outperforms classical solutions (supervised and unsupervised), DNN and recently-introduced methods [52, 74] in the same conditions. In Table 4, it can be seen that the performance of the proposed method in the speaker-independent case (68.89%) is better than that of the state-of-the-art results compared to 57.67% of [52] and 30.67% of [74]. Also, the method in speaker-dependent case (77.67%) performs better compared to 37.67% of [74] and 66.00% of [52] in term of accuracy, and the

Table 4 Three-folds folding case: Speaker-independent and speaker-dependent emotion recognition rate for the proposed pQPSO compared to classical and recently-used methods

Method	Three-folds Classification	
	Speaker-Independent (WAR)	Speaker-Dependent (WAR)
PCA	63.67%	73.00%
PPCA	63.67%	74.00%
FA	66.00%	74.33%
DNN	66.67%	77.00%
Yang et al., 2017 [74]	30.67%	37.67%
LDA	49.00%	63.67%
Pohjalainen et al., 2014 [52]	57.67%	66.00%
pQPSO (proposed)	**68.89%**	**77.67%**

best, compared to deep neural network classifier with accuracy rates, 66.67% and 77.00% in SI and SD cases, respectively.

In this comparison, as illustrated in Fig. 8, the recognition rate of the pQPSO in different iterations were significantly improved in both speaker-dependent and speaker-independent cases. Again, it is clear that for the same amount of training speech, a system trained on many speakers and tested on new speakers (i.e. speaker-independent recognition) has worse performance compared to the system trained on the speaker using it, namely speaker-dependent [20]. It should be noted that in this analysis the best results for SI and SD cases (73.45% and 78.49%, respectively), as illustrated in Fig. 8, are yielded after 100 iterations carried out on the first fold of the three-fold classification.

6.2 Second classification results

For another experiment, since LOSO folding strategy along with weighted average recall (WAR) is much more popular than speaker-dependent one and many new state of the art works evaluate their results according to it, then the algorithm accuracy and performance will be measured as follows.

Since the pQPSO initial random population makes different results for each execution, then all the executions have been repeated 10 times with 10 different random number generator (RNG) seeds, and the results for 10 runs are averaged. Also, WAR which is the average recognition rate of individual classes weighted by the class prior probability [81], has been adopted as recognition accuracy as follows,

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

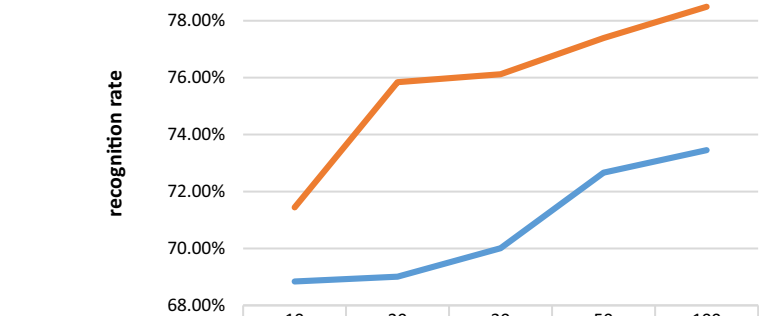
$$Weight_i = \frac{TP_i + FN_i}{N} \tag{3}$$

$$WAR = \sum_{i=1}^{M} Weight_i \times Recall_i \tag{4}$$

where M is the emotions' number, TP_i and FN_i are the numbers of true positive and false negative instances respectively for emotion i and N is the total number of instances from all emotions.

Because of the popularity of the LOSO folding strategy, the performance of the proposed method was measured only based on LOSO for EMO-DB, SAVEE, and IEMOCAP databases. In this manner, we took the test sets corresponding to each speaker. The training sets have been composed of 80% of the remaining speakers' utterances and the validation sets have been composed of another 20%. Therefore, we choose the 10-fold cross-validation strategy related to ten different speakers in EMO-DB and IEMOCAP databases and four-fold cross-validation strategy for SAVEE database to average over all possible choices of the test set.

A summary of the classification accuracy on EMO-DB, SAVEE, and IEMOCAP has been illustrated in Tables 5, 6 and 7 and Figs. 9, 10 and 11, respectively. According to Table 5, the



	10 iterations	20 iterations	30 iterations	50 iterations	100 iterations
speaker independent	68.84%	69.01%	70.01%	72.66%	73.45%
speaker dependent	71.44%	75.84%	76.12%	77.39%	78.49%

Fig. 8 Speaker-dependent and speaker-independent emotion recognition rate for pQPSO in different iterations (generations) for the first fold of the three-fold classification case

accuracy on EMO-DB in LOSO case increases significantly to 82.82%, obviously outperforms the previous works, 59.66% by [74], 68.49% by [52], and obtaining a 16.66%, 16.39% and 14.37% absolute improvement over the PCA, PPCA and FA methods, and also outperforms the recently-published state of the art works (Table 5).

Accuracies on SAVEE and IEMOCAP for the LOSO scheme are demonstrated in Tables 6 and 7 respectively. The tables compare results for several baseline methods and many recently-published works. The results show that accuracy on SAVEE improved by 9% over PCA, PPCA and FA cases and accuracy on IEMOCAP approximately improved by 20% over competing methods. Interestingly, the relative improvement in the results obtained with the proposed system on IEOMCAP is better than that on EMO-DB and SAVEE compared to the state of the art works.

Table 5 LOSO folding case: emotion recognition rate for the proposed pQPSO compared to classical and recently-published methods on EMO-DB

Method	WAR	Method	WAR
Yang et al., 2017 [74]	59.66%	Bashirpour et al., 2016 [3]	76.60%
PCA	66.16%	Luengo et al., 2010 [40]	78.30%
PPCA	66.43%	Zao et al., 2014 [79]	80.10%
LDA	67.36%	Bhargava et al., 2012 [4]	80.60%
DNN	68.26%	Badshah et al., 2019 [2]	80.79%
FA	68.45%	Zhang et al., 2013 [80]	80.85%
Pohjalainen et al., 2014 [52]	68.49%	Sun et al., 2015 [62]	81.50%
Sidorov et al., 2016 [57]	72.00%	Xu et al., 2015 [73]	81.80%
Yüncü et al., 2014 [77]	72.30%	Mak et al., 2016 [41]	81.86%
Khan et al., 2017 [33]	72.34%	Stuhlsatz et al., 2011 [61]	81.90%
Sinith et al., 2015 [58]	73.75%	Wen et al., 2017 [71]	82.32%
Deb et al., 2017 [12]	73.90%	Lotfidereshgi et al., 2017 [39]	82.35%
Deb et al., 2016 [11]	74.40%	Sun et al., 2017 [63]	82.40%
Kadiri et al., 2015 [30]	75.22%	Kalinli et al., 2016 [31]	82.70%
Shirani et al., 2016 [56]	76.12%	**pQPSO (proposed)**	**82.82%**

Table 6 LOSO folding case: emotion recognition rate for the proposed pQPSO compared to classical and recently-published methods on SAVEE

Method	WAR
Papakostas et al., 2014 [50]	44.00%
Liu et al., 2018 [38]	44.18%
Noroozi et al., 2017 [48]	45.51%
Vasquez-Correa et al., 2016 [67]	47.30%
LDA	50.49%
DNN	51.01%
FA	51.25%
PPCA	51.46%
Sun et al., 2017 [63]	51.46%
PCA	51.47%
Wen et al., 2017 [71]	53.60%
Tzinis et al., 2018 [66]	54.00%
Sinith et al., 2015 [58]	57.50%
Sun et al., 2015 [62]	58.76%
pQPSO (proposed)	**60.79%**

6.3 Comparison between different versions of QPSO method

To prove the supremacy of the proposed pQPSO algorithm, another experiment has been done. We have compared the proposed pQPSO with both standard QPSO [64] and a modified version of QPSO named wQPSO [72]. These algorithms have been executed in the same conditions for both three-folds and LOSO folding cases. The comparison between different folding strategies for 100 iterations running is illustrated in Table 8. The results show that the proposed pQPSO algorithm can achieve the best accuracy, 68.89%, and 77.67%, compared to 63.22% and 72.33% by wQPSO, and 67.78% and 75% by standard QPSO in both speaker-dependent and speaker-independent cases of three-folds folding for EMO-DB. Also, the proposed pQPSO can achieve the best accuracy, 82.82% compared to 81.69% and 79.91% by wQPSO and standard QPSO in LOSO folding case for EMO-DB and 58.96% and 59.58% by wQPSO and standard QPSO for SAVEE and 71.7% and 72.52% by wQPSO and standard QPSO respectively for IEMOCAP database.

Table 7 LOSO folding case: emotion recognition rate for the proposed pQPSO compared to classical and recently-published methods on IEMOCAP

Method	WAR
Ghosh et al., 2016 [19]	52.82%
PPCA	53.72%
PCA	54.12%
LDA	54.22%
DNN	61.59%
FA	61.64%
Li et al., 2015 [36]	63.20%
Tzinis et al., 2017 [65]	64.16%
Shirani et al., 2016 [56]	65.20%
Satt et al., 2017 [53]	68.80%
Li et al., 2018 [37]	71.75%
pQPSO (proposed)	**74.80%**

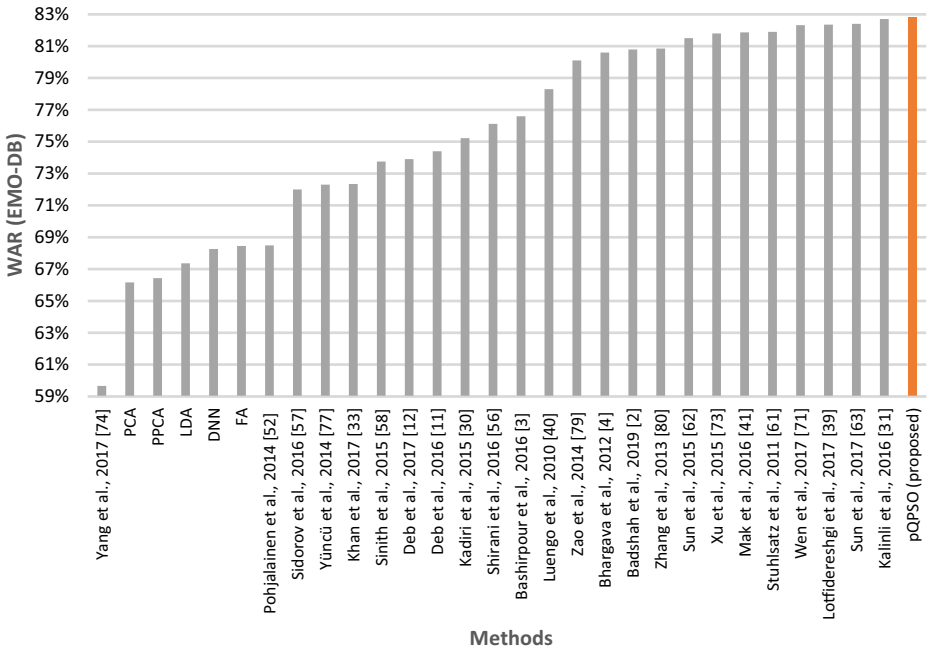


Fig. 9 LOSO folding case: emotion recognition rate compared to classical and recently-published methods on EMO-DB

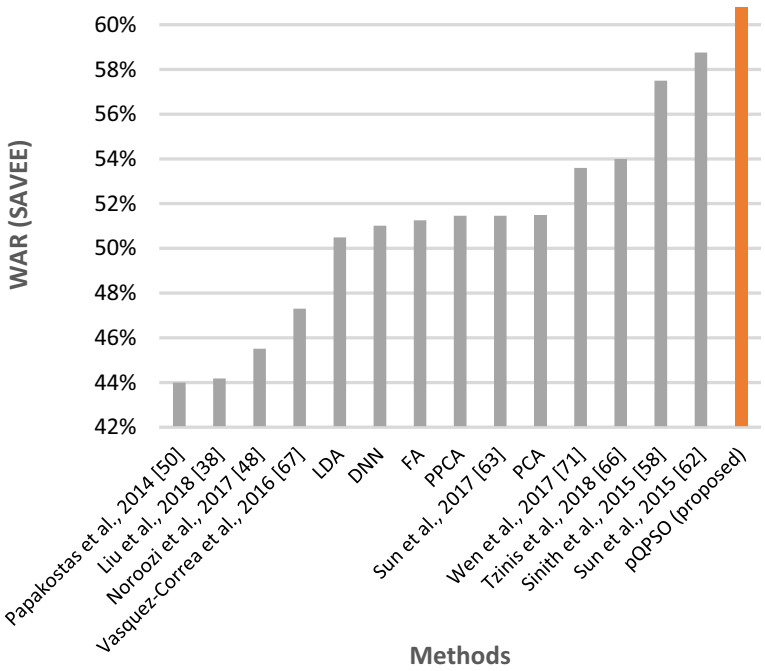


Fig. 10 LOSO folding case: emotion recognition rate compared to classical and recently-published methods on SAVEE

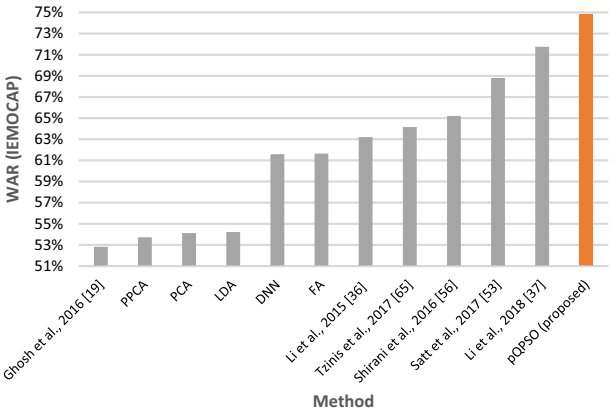


Fig. 11 LOSO folding case: emotion recognition rate compared to classical and recently-published methods on IEMOCAP

6.4 Per emotion performance

To illustrate the proposed algorithm performance and further investigate the recognition accuracy on each emotion separately, the confusion matrix corresponding to the LOSO folding strategy of EMO-DB, SAVEE, and IEMOCAP, have been illustrated in Figs. 12, 13, 14, 15, 16, 17, 18, 19 and 20 for pQPSO, wQPSO and standard QPSO methods. Figures 12, 13, 14, 15, 16 and 17 shows that for the case of EMO-DB and SAVEE datasets, "anger" is identified with the highest accuracy in all pQPSO, wQPSO and standard QPSO cases. However, Figs. 18, 19 and 20 indicates that "neutral" emotion is distinguished with the highest accuracy for IEMOCAP and pQPSO, wQPSO and standard QPSO algorithms.

6.5 Computational complexity

The proposed pQPSO method, like many metaheuristic algorithms, suffers from high computational complexity and low convergence speed in its iterative process, especially in high-dimensional spaces. It needs many iterations to converge to the optimum and also like PSO, it may be trapped in a local optimum.

Table 9 shows the computational complexity in terms of the real-time factor (RTF) which is meant to show the corresponding classification time of the proposed pQPSO algorithm in the test stage for three different databases and also the time needed to train the whole databases. The RTF (CPU Time divided by Audio Time) is a measuring factor

Table 8 Three-folds and LOSO folding: emotion recognition rate for standard QPSO and wQPSO compared to the proposed pQPSO

	Three-folds folding (WAR on EMO-DB)		LOSO folding (WAR)		
Method	Speaker-Independent	Speaker-Dependent	EMO-DB	SAVEE	IEMOCAP
Standard QPSO [64]	67.78%	75.00%	79.91%	59.58%	72.52%
wQPSO [72]	63.22%	72.33%	81.69%	58.96%	71.70%
pQPSO (proposed)	**68.89%**	**77.67%**	**82.82%**	**60.79%**	**74.80%**

Actual Class \ Predicted Class	F	D	J	B	N	S	A
Fear	66.66	0	7.24	2.89	10.14	04.34	08.69
Disgust	0	89.13	2.17	2.17	2.17	0	4.34
Joy	4.22	1.40	59.15	0	0	0	35.21
Bore	1.23	6.17	0	70.37	13.58	8.64	0
Neutral	0	2.53	0	8.86	87.34	1.26	0
Sadness	0	0	0	9.67	1.61	88.7	0
Anger	1.57	0	4.72	0	1.57	0	92.12

Fig. 12 Confusion matrix of standard QPSO on EMO-DB

for the speed of an SER system. When the RTF factor is smaller than one, the system will be real-time. It calculates the computational complexity of the emotion classification using the following equation,

$$\mathrm{RTF} = \frac{Time\ elapsed\ to\ classify\ an\ input\ signal}{Duration\ of\ the\ input\ signal} \tag{5}$$

All the experiments have been done on a system with Intel Core i3–4160 CPU at 3.60GHz and 8 GB RAM.

Actual Class \ Predicted Class	F	D	J	B	N	S	A
Fear	63.76	0	10.14	2.89	11.59	5.79	5.79
Disgust	2.17	91.30	2.17	0	0	2.17	2.17
Joy	5.63	1.4	60.56	0	0	0	32.39
Bore	0	2.46	0	77.77	12.34	7.4	0
Neutral	0	1.26	0	7.59	91.13	0	0
Sadness	0	0	0	11.29	0	88.70	0
Anger	2.36	0	4.72	0	0	0	92.91

Fig. 13 Confusion matrix of wQPSO on EMO-DB

Actual Class \ Predicted Class	F	D	J	B	N	S	A
Fear	72.46	1.44	4.34	0	10.14	4.34	7.24
Disgust	4.34	89.13	0	0	0	2.17	4.34
Joy	5.04	10.01	63.38	0	1.4	0	20.16
Bore	0	4.93	0	80.24	9.87	4.93	0
Neutral	1.26	2.53	0	5.06	91.13	0	0
Sadness	0	0	0	8.06	3.22	88.7	0
Anger	0.78	0	6.29	0	0	0	92.91

Fig. 14 Confusion matrix of pQPSO on EMO-DB

By the results, it can be concluded that due to a large number of experiments and low convergence of the proposed method, the training time is high; however, the low value of RTF in testing time indicates that the proposed algorithm response time is smaller than signal duration and then it is fast enough for real-time applications.

7 Discussion

The results of the proposed method show that the selection of simple features and effective dimensionality reduction method improves speech recognition accuracy. Also,

Actual Class \ Predicted Class	a	d	f	h	n	sa	su
a	80	6.66	5	1.66	3.33	1.66	1.66
d	16.66	51.66	0	8.33	15	5	3.33
f	16.66	5	35	8.33	1.66	1.66	31.66
h	20	3.33	3.33	46.66	0	6.66	20
n	7.5	15	0.83	0.83	68.33	7.5	0
sa	11.66	11.66	1.66	8.33	13.33	48.33	5
su	1.66	5	8.33	5	1.66	0	78.33

Fig. 15 Confusion matrix of standard QPSO on SAVEE

Actual Class \ Predicted Class	a	d	f	h	n	sa	Su
a	76.66	5	5	5	3.33	3.33	1.66
d	15	61.66	0	5	16.66	1.66	0
f	13.33	8.33	33.33	5	5	3.33	31.66
h	18.33	10	3.33	45	0	3.33	20
n	5.83	17.5	0.83	0	70	5	0.83
sa	18.33	18.33	1.66	5	10	43.33	3.33
su	3.33	5	11.66	6.66	1.66	0	71.66

Fig. 16 Confusion matrix of wQPSO on SAVEE

the presence of first and second-order derivatives of extracted features from speech and glottal signals eliminates existing incompatibilities between training and test data [28, 68]. Likewise, the strong correlation between glottal waveform features and the speaking style of an individual led to the use of features extracted from glottal waveform signals [45], which resulted in a more accurate speech emotions recognition system. Furthermore, the definition of the objective function in the proposed method, based on the

Actual Class \ Predicted Class	a	d	f	h	n	sa	su
a	90.00	0	5	0	3.33	0	1.67
d	10	68.33	1.67	1.67	18.33	0	0
f	31.67	6.67	35	3.33	5	5	13.33
h	28.33	6.67	3.33	50	5	0	6.67
n	7.5	19.17	0.83	0	67.5	4.17	0.83
sa	10	25	3.33	0	18.33	41.67	1.67
su	16.67	1.67	10	0	1.67	1.67	68.33

Fig. 17 Confusion matrix of pQPSO on SAVEE

Actual Class \ Predicted Class	Anger	Happiness	Neutral	Sadness
Anger	69.37	7.52	11.64	11.45
Happiness	6.273	67.44	12.2	14.08
Neutral	4.95	5.93	84.06	5.04
Sadness	3.27	10.6	18.45	67.67

Fig. 18 Confusion matrix of standard QPSO on IEMOCAP

accuracy of emotion classification in the development set, improves the results on the test set. Another advantage of the proposed method is the simultaneous estimation of the projection matrix and GMM parameters.

Despite the strengths of the proposed algorithm, the pQPSO has an iterative process, and consumes a large amount of memory, like other iterative algorithms, with high complexity and low convergence speed. In the training phase, parallel computers or GPUs can be used to solve this problem. Although the proposed algorithm has a high learning time, its response time is less than the signal duration, in other words, it is fast enough for real-time processing and can be useful in real-time applications.

This paper aims to provide a simple and general framework for the better recognition of speech emotions with a few parameters. However, there is still a lot of open issues in speech emotion recognition to be investigated in future researches. Combining more features like predictor features [55] and performing cross-corpus emotion recognition experiments on IEMOCAP, EMO-DB and SAVEE datasets are considered in future studies. Also, MiGSA, a new simulated annealing algorithm [43] is another proposed

Actual Class \ Predicted Class	Anger	Happiness	Neutral	Sadness
Anger	70.08	3.93	12.15	13.83
Happiness	7.27	65.98	11.47	15.25
Neutral	5.14	6.28	82.35	6.21
Sadness	3.75	10.86	18.39	66.97

Fig. 19 Confusion matrix of wQPSO on IEMOCAP

Actual Class \ Predicted Class	Anger	Happiness	Neutral	Sadness
Anger	70.40	4.97	12.71	11.92
Happiness	4.13	74.20	12.88	8.79
Neutral	5.14	6.07	84.39	4.40
Sadness	6.24	13.21	19.25	61.30

Fig. 20 Confusion matrix of pQPSO on IEMOCAP

optimization algorithm by the authors which will be a research direction for future work instead of the pQPSO algorithm for transformation matrix estimation. In the classification stage of the system, it seems that employing dynamic acoustic models like hidden Markov models or time-inhomogeneous hidden Bernoulli models [29] will be more suitable than static acoustic models like Gaussian mixture models for speech emotion recognition task.

8 Conclusion

In this paper, a modified version of the QPSO algorithm called pQPSO is proposed to solve the dimensionality reduction problem in speech recognition systems. The dimensionality of extracted feature vectors of speech and glottal signals, prosodic features and their first and second-order derivatives are reduced by the use of the transformation matrix estimated by the pQPSO algorithm. Also, a proper number of cepstral coefficients and principal components have been estimated experimentally for further investigations. The results of applying our method on large emotional datasets such as EMO-DB, SAVEE, and IEMOCAP show that in terms of accuracy, the proposed pQPSO algorithm outperforms standard QPSO algorithms, wQPSO, classical dimensionality reduction methods, deep neural networks, and the state-of-the-art methods on the same datasets. In addition to optimizing the GMM parameters and the transformation matrix for dimensionality reduction, other applications of the pQPSO algorithm include optimizing the MFCC filter bank parameters, optimizing the classifier parameters, and finding more features related to emotion.

Table 9 The computational complexity of the proposed pQPSO

Dataset	Training time	Emotion classification RTF (xRT)
EMO-DB	27,860 s	0.6061
SAVEE	7296 s	0.3961
IEMOCAP	140,290 s	0.5761

Despite these applications of the pQPSO, high memory consumption and high computational complexity are the limitations of the proposed method to be investigated in future studies. Also, the authors are currently working on powerful classifiers such as the deep extreme learning machines and elliptical basis function networks, and plan to extend the proposed method to these classifiers.

Acknowledgments We hereby express our gratitude to Abbas Neekabadi for providing us with some source codes.

References

1. Albornoz EM, Milone DH, Rufiner HL (2017) Feature extraction based on bio-inspired model for robust emotion recognition. Soft Comput 21(17):5145–5158
2. Badshah AM, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee MY, Kwon S, Baik SW (2019) Deep features-based speech emotion recognition for smart affective services. Multimed Tools Appl 78(5):5571–5589
3. Bashirpour M, Geravanchizadeh M (2016) Speech emotion recognition based on power normalized cepstral coefficients in noisy conditions. Iranian Journal of Electrical and Electronic Engineering 12(3):197–205
4. Bhargava M and Polzehl T (2012) Improving automatic emotion recognition from speech using rhythm and temporal feature, *Proc. International Conference on Emerging Computation and Information Technologies*
5. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, and Weiss B (2005) A database of German emotional speech. Ninth European Conference on Speech Communication and Technology
6. Buscicchio CA, Górecki P and Caponetti L (2006) Speech emotion recognition using spiking neural networks. *International Symposium on Methodologies for Intelligent Systems*. Springer Berlin Heidelberg
7. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335
8. Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: features and classification models. Digital signal processing 22(6):1154–1160
9. Cho Y-H, Park K-S, and Pak RJ (2007) Speech emotion pattern recognition agent in mobile communication environment using fuzzy-SVM. *Fuzzy information and engineering*. Springer Berlin Heidelberg, 419–430
10. Darekar RV, Dhande AP (2018) Emotion recognition from Marathi speech database using adaptive artificial neural network. Biologically Inspired Cognitive Architectures 23:35–42
11. Deb S, and Dandapat S (2016) Emotion classification using residual sinusoidal peak amplitude. *2016 International Conference on Signal Processing and Communications (SPCOM)*. IEEE
12. Deb S, and Dandapat S (2017) Exploration of phase information for speech emotion classification. *2017 Twenty-third National Conference on Communications (NCC)*. IEEE
13. Degottex G, Kane J, Drugman T, Raitio T, and Scherer S (2014) COVAREP—A collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE
14. Deng J, Zhang Z, Eyben F, Schuller B (2014) Autoencoder-based unsupervised domain adaptation for speech emotion recognition. IEEE Signal Processing Letters 21(9):1068–1072
15. Duda RO, Hart PE, and Stork DG. (2001) *Pattern classification, 2nd Ed*. John Wiley & Sons
16. Gangeh MJ, Fewzee P, Ghodsi A, Kamel MS, Karray F (2014) Multiview supervised dictionary learning in speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(6):1056–1068
17. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S (2012) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput & Applic 21(8):2115–2126
18. Gharavian D, Bejani M, Sheikhan M (2017) Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks. Multimed Tools Appl 76(2):2331–2352
19. Ghosh S, Laksana E, Morency L-P, and Scherer S (2016) Representation Learning for Speech Emotion Recognition. Interspeech
20. Grimm M, Kroschel K, Mower E, Narayanan S (2007) Primitives-based evaluation and estimation of emotions in speech. Speech Comm 49(10):787–800

21. Haq S, and Jackson PJB. (2011) Multimodal emotion recognition. *Machine audition: principles, algorithms and systems*. IGI Global, 398–423
22. Yogesh CK, Hariharan M, Ngadiran R, Adom AH, Yaacob S, Berkai C, Polat K (2017) A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. Expert Syst Appl 69:149–158
23. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87(4):1738–1752
24. Huang Y, Wu A, Zhang G, Li Y (2015) Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition. IET Signal Processing 9(4):341–348
25. Huang Z-w, Xue W-t, Mao Q-r (2015) Speech emotion recognition with unsupervised feature learning. Frontiers of Information Technology & Electronic Engineering 16(5):358–366
26. Huang Z, Xue W, Mao Q, Zhan Y (2017) Unsupervised domain adaptation for speech emotion recognition using PCANet. Multimed Tools Appl 76(5):6785–6799
27. Idris I and Salam MS (2016) Improved Speech Emotion Classification from Spectral Coefficient Optimization. *Advances in Machine Learning and Signal Processing*. Springer International Publishing, 247–257
28. Junqua J-C, and Haton J-P (2012) *Robustness in automatic speech recognition: Fundamentals and applications*. Vol. 341. Springer Science & Business Media
29. Kabudian J, Mehdi Homayounpour M, and Mohammad Ahadi S (2008) Time-inhomogeneous hidden Bernoulli model: An alternative to hidden Markov model for automatic speech recognition. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE
30. Kadiri SR, Gangamohan P, Gangashetty SV, and Yegnanarayana B (2015) Analysis of excitation source features of speech for emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*
31. Kalinli O (2016) Analysis of Multi-Lingual Emotion Recognition Using Auditory Attention Features. INTERSPEECH
32. Keyvanrad MA, and Homayounpour MM. (2014) A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet). *arXiv preprint arXiv:1408.3264*
33. Khan A and Roy UK (2017) Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. *2017 international conference on wireless communications, signal processing and networking* (WiSPNET). IEEE
34. Kim EH, Hyun KH, Kim SH, Kwak YK (2009) Improved emotion recognition with a novel speaker-independent feature. IEEE/ASME Transactions on Mechatronics 14(3):317–325
35. Li X, Li X, Zheng X, Zhang D (2010) EMD-TEO Based speech emotion recognition. *Life System Modeling and Intelligent Computing*. Springer Berlin Heidelberg. 180–189
36. Li Y, Chao L, Liu Y, Bao W, and Tao J (2015) From simulated speech to natural speech, what are the robust features for emotion recognition?. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE
37. Li P, Song Y, McLoughlin I, Guo W, and Dai L (2018) An Attention Pooling based Representation Learning Method for Speech Emotion Recognition. *Proc. Interspeech* (2018): 3087–3091
38. Liu Z-T, Xie Q, Wu M, Cao W-H, Mei Y, Mao J-W (2018) Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing 309:145–156
39. Lotfidereshgi R, and Gournay P (2017) Biologically inspired speech emotion recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE
40. Luengo I, Navas E, Hernáez I (2010) Feature analysis and evaluation for automatic emotion identification in speech. IEEE Transactions on Multimedia 12(6):490–501
41. Mak MW (2016) Feature Selection and Nuisance Attribute Projection for Speech Emotion Recognition, Technical Report and Lecture Note Series, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
42. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia 16(8):2203–2213
43. Mirhosseini SH, Yarmohamadi H, and Kabudian J (2014) MiGSA: A new simulated annealing algorithm with mixture distribution as generating function. *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE
44. Mistry K, Zhang L, Neoh SC, Lim CP, Fielding B (2016) A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. IEEE Transactions on Cybernetics 47(6):1496–1509
45. Moore E II, Clements MA, Peifer JW, Weisser L (2007) Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans Biomed Eng 55(1):96–107
46. Muthusamy H, Polat K, Yaacob S (2015) Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals. PLoS One 10(3):e0120344

47. Neekabadi A, and Kabudian SJ. (2018) A New Quantum-PSO Metaheuristic and Its Application to ARMA Modeling of Speech Spectrum. *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE
48. Noroozi F, Sapiński T, Kamińska D, Anbarjafari G (2017) Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology 20(2):239–246
49. Pant M, Thangaraj R, and Abraham A (2008) A new quantum behaved particle swarm optimization. *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. ACM
50. Papakostas M, Spyrou E, Giannakopoulos T, Siantikos G, Sgouropoulos D, Mylonas P, Makedon F (2017) Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. Computation 5(2):26
51. Park J-S, Kim J-H, Yung-Hwan O (2009) Feature vector classification based speech emotion recognition for service robots. IEEE Trans Consum Electron 55(3):1590–1596
52. Pohjalainen J and Alku P (2014) Multi-scale modulation filtering in automatic detection of emotions in telephone speech. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE
53. Satt A, Rozenberg S, and Hoory R (2017) Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. INTERSPEECH
54. Sheikhan M, Bejani M, Gharavian D (2013) Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. Neural Comput & Applic 23(1):215–227
55. Shekofteh Y, Kabudian J, Goodarzi MM, Rezaei IS (2012) Confidence measure improvement using useful predictor features and support vector machines. *20th Iranian Conference on Electrical Engineering (ICEE2012)*. IEEE
56. Shirani A, and Nilchi ARN (2016) Speech Emotion Recognition based on SVM as Both Feature Selector and Classifier. International Journal of Image, Graphics & Signal Processing 8, 4
57. Sidorov M, Minker W, Semenkin ES (2016) Speech-based emotion recognition and static speaker representation. Journal of the Siberian Federal University The series Mathematics and Physics 9(4):518–523
58. Sinith MS, Aswathi E, Deepa TM, Shameema CP and Rajan S (2015) Emotion recognition from audio signals using Support Vector Machine. *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE
59. Song P, Jin Y, Cheng Z, Zhao L (2015) Speech emotion recognition method based on hidden factor analysis. Electron Lett 51(1):112–114
60. Stewart GW (1998) *Matrix Algorithms: Volume 1: Basic Decompositions*. Vol. 1. SIAM
61. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, and Schuller B. (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5688–5691. IEEE
62. Sun Y, Wen G (2015) Emotion recognition using semi-supervised feature selection with speaker normalization. International Journal of Speech Technology 18(3):317–331
63. Sun Y, Wen G (2017) Ensemble softmax regression model for speech emotion recognition. Multimed Tools Appl 76(6):8305–8328
64. Sun J, Lai C-H, and Wu X-J. (2011) *Particle swarm optimisation: classical and quantum perspectives*. CRC Press
65. Tzinis E, and Potamianos A (2017) Segment-based speech emotion recognition using recurrent neural networks. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE
66. Tzinis E, Paraskevopoulos G, Baziotis C, and Potamianos A. (2018) Integrating Recurrence Dynamics for Speech Emotion Recognition. Proc. Interspeech (2018): 927–931
67. Vasquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Vargas-Bonilla JF and Noeth E (2016) Wavelet-based time-frequency representations for automatic recognition of emotions from speech. *Speech Communication; 12. ITG Symposium. VDE*
68. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Comm 48(9):1162–1181
69. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using Fourier parameters. IEEE Trans Affect Comput 6(1):69–75
70. Wang S-H, Phillips P, Dong Z-C, Zhang Y-D (2018) Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. Neurocomputing 272:668–676
71. Wen G, Li H, Huang J, Li D, and Xun E (2017) Random deep belief networks for recognizing emotions from speech signals. Computational intelligence and neuroscience 2017
72. Xi M, Sun J, Xu W (2008) An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position. Appl Math Comput 205(2):751–759

73. Xu X, Deng J, Zheng W, Zhao L, and Schuller B (2015) Dimensionality reduction for speech emotion features by multiscale kernels. *Sixteenth Annual Conference of the International Speech Communication Association*
74. Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W, Sturge-Apple M (2017) Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. International Journal of Speech Technology 20(1):27–41
75. Yapanel UH, Hansen JHL (2008) A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. Speech Comm 50(2):142–152
76. Yogesh CK, Hariharan M, Ngadiran R, Adom AH, Yaacob S, Polat K (2017) Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech. Appl Soft Comput 56:217–232
77. Yüncü E, Hacihabiboglu H, and Bozsahin C (2014) Automatic speech emotion recognition using auditory models with binary decision tree and svm. *2014 22nd International Conference on Pattern Recognition. IEEE*
78. Zaidan NA, and Salam MS (2016) MFCC Global Features Selection in Improving Speech Emotion Recognition Rate. *Advances in Machine Learning and Signal Processing*. Springer International Publishing, 141–153
79. Zao L, Cavalcante D, Coelho R (2014) Time-frequency feature and AMS-GMM mask for acoustic emotion classification. IEEE Signal Processing Letters 21(5):620–624
80. Zhang S, Zhao X, and Lei B (2013) Speech emotion recognition using an enhanced kernel isomap for human-robot interaction, Int J Adv Robot Syst, 10
81. Zhang S, Zhang S, Huang T, Gao W (2017) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Transactions on Multimedia 20(6):1576–1590
82. Zhao Z, Zhao Y, Bao Z, Wang H, Zhang Z, and Li C (2018) Deep Spectrum Feature Representations for Speech Emotion Recognition. *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data. ACM*
83. Zheng W, Xin M, Wang X, Wang B (2014) A novel speech emotion recognition method via incomplete sparse least square regression. IEEE Signal Processing Letters 21(5):569–572
84. Zong Y, Zheng W, Zhang T, Huang X (2016) Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. IEEE Signal Processing Letters 23(5):585–589