Speech Emotion Recognition Using Enhanced Cat Swarm Optimization Algorithm

Amarjeet Singh
Coordinator of Department Pyramid College of Business & Technology - Phagwara

ABSTRACT

Human interactions involve emotional cues which can be used to interpret the emotion expressed by the speaker. As the vocal emotions vary from one speaker to another, there is a chance of misinterpretation. So as to determine the emotion expressed by the speaker, speech emotion recognizer can be utilized. It is known that speech expresses the emotional states of humans along with the syntax and semantic content of linguistic sentences. Therefore, human emotion recognition using speech signaling is possible. Speech emotion recognition is a crucial and challenging task in which the feature extraction plays a prominent role in its performance. Determining emotional states in speech signals is a very challenging area for many reasons. The first issue of all speech emotion systems is the selection of the best features, which is powerful enough to distinguish various emotions.

Keywords: Speech Recognition, Human Interactions involves.

I. INTRODUCTION

The speech signal consists of linguistic information and also paralinguistic one such as emotion. The modern automatic speech recognition systems have achieved high performance in neutral style speech recognition [1-15]. The acoustic and prosodic features of speech are affected by emotions and speaking styles as well as speaker characteristics and linguistic features. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, and improving the voice-based man–machine interactions [16-25]. Man–machine interaction is one of the key goals in developing automatic emotion recognition (AER) systems. The AER system is a key component in many applications such as spoken tutoring systems, medical-emergency domain to detect stress and pain, interactions with robots, computer games, call centers and developing man–machine interfaces for helping people [26-35]. In the field of multimedia contents management, it is used for emotional labeling and retrieval of the contents [36-42]. In computer game, identification of the emotional state of a player is used to assess the interest of the player. In the field of computer-based tutorial system, the learning rate of students can be improved by making the system take into account the emotional states of students. Generally, the emotion recognition system has three components: feature extraction unit, feature selection unit and emotion recognition unit [43-44]. However, the performance of the emotion recognition is still far from the expectation of researchers. In speech emotion recognition, there are mainly two difficulties that are how to find effective speech emotion features, and how to construct a suitable speech emotion recognition model [45]. Some factors such as the number and gender of speakers, dialect, age, language, and skills are the effective factors of emotion recognition accuracy. Extracting a limited, meaningful, and informative set of features is an important step in automatic recognition of emotions [46-55]. The irrelevant features reduce the correct classification rates. Feature fusion will resolve most of the issues, but it may lead to high dimension and redundancy of features, so it is vital to filter out the characteristic parameters of higher distinguish ability [56]. The interaction between features generated from the same audio source was rarely considered, which may produce redundant features and increase the computational costs. So as to solve these issues, many of the approaches have been proposed using Neural Network, Decision Tree, Support Vector Machine and K-Nearest Neighbor. But most of them were time consuming [57-68]. The selection of feature subsets in some methods relied only on the performance of the classifier. And in many of the approaches, the accuracy drops significantly for non-acted spontaneous emotional speech [69].

II. LITERATURE SURVEY

To increase the emotional recognition success and reduce workload with fewer features Ozseven [71] has proposed a new approach. A new Statistical feature selection approach was proposed in terms of changes
in emotion in sound features. The success rate has increased in the EMOVO data set, but this increase was 2.37% lower than PCA. To create prominence annotations, extract prominence features and apply them in speech emotion recognition Jing et al [72] have proposed an approach. The annotation rules were initially designed and for verifying the annotation consistency of annotated samples, a consistency assessment algorithm has been presented which has been used in solving the issue of lack of quality estimation method for speech annotation. Finally, for assisting the recognition of emotions, novel features related to prominence has been presented and the correlation among the proposed features and typical emotions was analyzed. The annotated prominence features were so widely scattered, making it hard to find the prominence trends through visual inspection [73].

To learn local correlation and global contextual information from raw audio clips and log-mel spectrogram, Zhao et al [15] have proposed an approach. Two convolution neural network and long short-term memory (RNN LSTM) networks, one 1D CNN LSTM network and one 2D CNN LSTM network, were constructed in the proposed approach. The two networks have the similar architecture, both consisting of four local feature learning blocks (LFLBs) and one long short-term memory (LSTM) layer. The designed networks recognize the emotion cannot be explained in more detail, meaning the “black box” of these networks have not been uncovered. It has not been able to acquire higher accuracy in speech emotion recognition [74]. A drawback to this method is that the proposed approach has not been implemented in any other speech recognition task. This has lowered the trustworthiness of the presented approach.

A parallel convolution recurrent neural network with special features was proposed by Jiang et al [75] for speech emotion recognition. The main goal of the proposed approach has been to extract the best emotional features from speech signals. Though this method's performance has been superior to other methods, the accuracy has been nearly as same as other existing approaches.

III. OVERVIEW OF PROPOSED METHODOLOGY

The main objective of the proposed method is to select the features from the speech signal and to classify the selected feature based on emotion. Feature selection is one of the important steps in determining emotions in speech. Ignoring the connection between the features selected from the same audio clip results in repetition of features and increases the computational cost. These issues can be handled by selecting optimal features. In this work, Cat Swarm Optimization (CSO) algorithm is presented for optimal feature selection. And to improve the searching ability, Opposition Based Learning (OBL) is included in CSO. The speech signal is first processed with the CSO algorithm in which the features are extracted. And for optimal feature selection, oppositional based learning is used. The optimal features thus obtained are provided as input to the classifier for speech emotion recognition. The SVNN classifier is used for classifying the emotions which are given as the input. Feature selection approach is utilized for feature extraction and classification approach is utilized for emotion recognition.

![Feature selection diagram](image_url)
Basic CSO

In the CSO algorithm, cats and the model of behaviors of cats are used to solve the optimization problems, i.e. Cats are used to portray the solution sets. In CSO, a decision has to be made on how many cats are to be used and then the cats are applied into CSO to solve the problems. Every cat has its own position composed of N dimensions, velocities for each dimension, a fitness value, which represents the accommodation of the cat to the fitness function, and a flag to identify whether the cat is in seeking mode or tracing mode. The final solution would be the best position in one of the cats because CSO keeps the best solution till it reaches the end of iterations.

Seeking Mode

This sub-model is used to model the situation of the cat, which is resting, looking around and seeking the next position to move to. In seeking mode, we define four essential factors: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC), self-position considering (SPC) and Mixture Ration (MR).

SMP is used to define the size of seeking memory for each cat, which indicates the points sought by the cat. The cat would pick a point from the memory pool according to the rules described later.

SRD declares the mutative ratio for the selected dimensions. In seeking mode, if a dimension is selected to mutate, the difference between the new value and the old one will not out of the range, which is defined by SRD.

CDC discloses how many dimensions will be varied. These factors are all playing important roles in the seeking mode.

SPC is a Boolean variable, which decides whether the point, where the cat is already standing, will be one of the candidates to move to. No matter the value of SPC is true or false; the value of SMP will not be influenced.

MR: To guarantee that the cats spend most of their time resting and observing, i.e., most of the time is spent in seeking mode, a term called mixture ratio (MR), which is a fraction of population allocated a very small value.

The working of seek mode can be described using 5 steps as follows:

1. Step1: Make k copies of the present position of cat, where k = SMP. If the value of SPC is true, let k = (SMP-1), then retain the present position as one of the candidates.

2. Step2: For each copy, according to CDC, randomly plus or minus SRD percents of the present values and replace the old ones.

3. Step3: Calculate the fitness values (FV) of all candidate points.

4. Step4: If all FV are not exactly equal, calculate the selecting probability of each candidate point by equation (1), otherwise set all the selecting probability of each candidate point be 1.

5. Step5: Randomly pick the point to move to from the candidate points, and replace the position of cat.

$$Q_j = \frac{|FV_j - FV_e|}{FV_{\text{max}} - FV_{\text{min}}} , \text{where } 0 < j < k$$

(1)

If the goal of the fitness function is to find the minimum solutions, \(FV_e = FV_{\text{max}}\), otherwise \(FV_e = FV_{\text{min}}\).

Tracing Mode

Tracing mode is the sub-model for modeling the case of the cat in tracing some targets. Once a cat goes into tracing mode, it moves according to its’ own velocities for every dimension. The action of tracing mode can be described in 3 steps as follows:

1. Step1: Update the velocities for every dimension \(w_{i,e}\) according to equation (2).

2. Step2: Check if the velocities are in the range of maximum velocity. In case the new velocity is over-range, set it be equal to the limit.

3. Step3: Update the position of cat \(y_{i,e}\) according to equation (3).

$$w_{i,e} = w_{i,e} + s_1 \times d_1 \times (y_{\text{best},e} - y_{i,e}) , \text{where } e = 1,2,\ldots,N$$

(2)

\(y_{\text{best},e}\) is the position of the cat, who has the best fitness value; \(y_{i,e}\) is the position of cat \(i\); \(d_1\) is a constant and \(s_1\) is a random value in the range of \([0,1]\).

$$y_{i,e} = y_{i,e} + w_{i,e}$$

(3)
For enhancing the population diversity of the CSO algorithm, the OBL method is used in CSO algorithm. According to OBL, if Y is the solution to a given problem; then, the opposite of Y would be the other candidate solution. Thus the chance of obtaining optimal solution is increased.

Let \( y \in [b, c] \), the opposite number \( y' \) is defined as

\[
y' = b + c - y
\]

The above definition can be extended to higher dimensions as follows:

Let \( Q(y_1, y_2, \ldots, y_m) \) be an m-dimensional vector, where \( y_j \in [b_j, c_j] \) and \( j=1,2,\ldots,n \). The opposite vector of \( Q \) is defined by \( Q'(y'_1, y'_2, \ldots, y'_m) \) where

\[
y'_j = b_j + c_j - y_j
\]

d. Termination:

Check the termination condition, if satisfied, terminate the program, otherwise repeated.

IV. RESULT AND DISCUSSION:

The proposed Enhanced ECSO based SVNN classifier is experimented and implemented on the MATLAB tool. The simulation is carried out on a Intel I3 processor powered PC that runs on Windows 7 OS.

Figure 2 shows the performance of the proposed ECSO-SVNN approach and existing approaches such as CSO-SVNN and PSO-SVNN in terms of accuracy. The accuracy of an approach has to be high for it to be considered as an effective approach. The average accuracy of ECSO-SVNN is 0.96, the average accuracy of CSO-SVNN is 0.89 and the average accuracy of PSO-SVNN is 0.91. Figure 3 shows the sensitivity performance of the proposed ECSO-SVNN approach and the sensitivity performance of current approaches such as CSO-SVNN and PSO-SVNN. The sensitivity of an approach must be high to be effective. The sensitivity of the proposed ECSO-SVNN approach is 0.74, the sensitivity of the CSO-SVNN approach is 0.69 and the sensitivity of the PSO-SVNN approach is 0.7. The sensitivity of the proposed approach is higher than the existing approaches. From the graph, it can be concluded that the proposed ECSO-SVNN approach performed better in terms of sensitivity than the existing CSO-SVNN and PSO-SVNN approaches.
V. CONCLUSION

A speech emotion recognizer is presented based on the optimal feature extraction. In the presented approach, Cat Swarm Optimizer has been utilized for feature selection and Opposition Based Learning has been utilized for better search. Using the proposed approach, only the necessary features are selected and extracted. And for classification purpose, SVNN classifier has been used in which the selected features are provided as the input. Based on the training, SVNN classifier classifies the emotions for the provided input features. The results have been simulated for the proposed approach and compared with two existing approaches. The performance on accuracy, sensitivity, specificity and recognition rate of the proposed approach has been simulated. From the result, it can be concluded that the proposed ECSO-SVNN approach has performed better than the compared existing approaches.

REFERENCES


static performance of multilevel inverter for single-phase grid connected photovoltaic modules. In 2009 Second International Conference on Emerging Trends in Engineering & Technology (pp. 691-697). IEEE.


