Application of Egyptian Vulture Optimization in Speech Emotion Recognition

Shreya Sahu[1], *Arpan Jain*[2], *Anupam Shukla*[3], *Ritu Tiwari*[4]

[1]ABV-Indian Institute of Information Technology and Management Gwalior
[2]ABV-Indian Institute of Information Technology and Management Gwalior
[3]ABV-Indian Institute of Information Technology and Management Gwalior
[4]ABV-Indian Institute of Information Technology and Management Gwalior

shreya.iiitm@gmail.com, arpan.jain1405@gmail.com, anupamshukla@iiitm.ac.in, tiwariritu2@gmail.com

Abstract

Recognition of emotions present in human speech is a task made complicated due to incomplete knowledge of the relevant features, dependency on language, dialect and even individuals and the ambiguous meaning of term emotion itself. This work aims to study the relevant features present in human speech and infer its emotional category, making use of machine learning models.

The prime objective of this work is to analyze the presented human speech and classify it into one of the seven emotional categories- happiness, sadness, anger, boredom, anxiety, disgust, and neutral. Proposed method extracts Mel Frequency Cepstral coefficients (MFCC), Chroma and Time Spectral features from Berlin EmoDB speech clips, and compares the results obtained from traditional classifiers to those with added nature inspired optimization technique.

Egyptian Vulture Optimization Algorithm (EVOA),Grey Wolf Optimization(GWO) and Moth Flame Optimization(MFO) were applied over the mentioned feature set along with RF, KNN and SVM, and it is inferred that EVOA significantly improves the performance metrics of mentioned algorithms over EMO-DB when applied alongside. Using GWO with SVM produced the highest classification accuracy 90.67%.

Keywords: Speech, Emotion, KNN, SVM, RF, Egyptian Vulture, EVOA, GWO, MFO

1. INTRODUCTION

Emotions are the portrayal of the internal conscious condition of a human personality. They stimulate discussions, help in understanding the perspectives other than just phonetic, and in increasing more about human conduct. People are perfect works of art in both communicating and deciphering emotions, because of our senses that have been feeding information into our cerebrum storage house, each millisecond right from the time we were born. Added to giving an unmistakable sign of the perspective of the speaker, emotions add depth and richness to the conversation. For several decades now, considering the headway emotions add to speech, researchers have been endeavoring to execute passionate information in machines additionally, with the essential intend to allow human-machine interaction to approach human-human conversation.

The principle goal of utilizing speech emotion recognition is to adjust the system response after recognizing dissatisfaction, inconvenience or distress in the speaker's voice [1]. Therapists may use it as a diagnostics tool [3], it can be used in systems where mental state of user plays a key role, for example automatic translation systems. It can be used in in-car board systems, where mental state of the driver would be provided to the system to ensure his safety [2]. It also has demonstrated enormous utility in mobile communication and call center applications. SER is especially valuable for applications which require characteristic man machine interaction, for example, web movies and computer tutorial applications where the reaction of those frameworks to the client relies upon the identified emotion. [2].

Getting the computer to comprehend emotions from speech is a challenging work. Human speech conveys a lot of information: phonetic-identified with spoken words, paralinguistic-identified with deeper context, i.e. non-verbal communication, tone, pitch, facial expressions. In addition, an utterance may correspond to various emotions, one being for each sub-sentence. Adding to the multifaceted nature, every language forces a noteworthy changeability in it's expression, and that every speaker gives his very own personal characteristics while communicating.

Complex basic variables like varieties of speaker and content, and ecological variations [4], [6] make SER challenging. Also, another of the fundamental challenge in SER is categorization and meaning of emotion itself. From quite a while, various disciplines have had distinctive meanings of the same. Most present SER depends on crafted by [7], as indicated by whom there are six essential all around perceived feelings happiness, sadness, surprise, fear, anger and disgust. Few databases likewise include 'neutral' as a category too. Also there exists an arrangement of emotion on a two-dimensional excitement valence plane [8].

Selection of relevant features from speech is one of the most crucial tasks in emotion recognition [6]. Over time this field has witnessed many algorithms, like Support Vector Machine (SVM) [15], Artificial Neural Network (ANN) [13], K Nearest Neighbours (KNN) [14], Hidden Markov Model (HMM) [11], Gaussian Mixture Model (GMM) [12]. Among these methods, SVM and HMM are the most widely used ones. The look for more strong classifiers that can function admirably with existing speech and spectral features is as yet going on, and with time the utilization of more conventional methodologies like HMM and GMM is gradually diminishing.

Nature Inspired Algorithms (NIA) are based on the inbuilt intelligent conduct and intellect of creatures shaped by nature, given the shape of algorithms. This work is based on a similar conduct of a bird, egyptian vulture, which imparts it's smartness in the way it hunts and procures food. More is mentioned in the upcoming sections.

10.21437/SLTU.2018-48

2. DATASET AND METHODOLOGY

2.1. Dataset

Three kinds of speech are observed. Natural speech is simply spontaneous speech where all emotions are real. Simulated or acted speech is speech expressed in a professionally deliberated manner. Finally, elicited speech is speech in which the emotions are induced. [16]
Proposed Speech Emotion Recognition model was trained upon Berlin Emotional Dataset (Berlin EMO-DB) [16]. which contained 535 utterances by 10 German actors, 5 male and 5 female, recorded in an artificially created environment sampled at 16kHz. An aggregate of ten distinct sentences were produced in seven emotions. Each emotional utterance is from one of the seven emotions - anger, boredom, disgust, fear, happy, neutral, and sad.

2.2. Methodology

Literature reveals numerous feature extraction techniques and strategies from speech signals. Few of the best include MFCC features, fundamental frequency, wavelet transform. This work makes use of statistical data obtained from Energy, spectral, MFCC and chroma features.

Feature Extraction has been carried out from an open source library for audio feature extraction [18]. Short term features are obtained by dividing audio signal into small frames and extracting local features from each frame. 34 short term features are extracted: Zero Crossing Rate (ZCR), Energy - 2 Features, Spectral - 5 Features, MFCC - 13 Features, Chroma - 13 Features. Five core data representatives are used- mean, median, standard deviation, maximum, and minimum.

2.2.1. Zero Crossing Rate (ZCR)

It is the rate of sign change of a signal, or the number of time the signal changes from negative to positive or vice versa in a time period (usually one second) [[19]].

2.2.2. Energy features

Energy- Sum of squares of acoustic signal values, normalized by respective frame lengths is called Energy.

Energy Entropy- Interpreted as a measure of abrupt change. Entropy of sub frames' normalized energies is called entropy of energy.

2.2.3. Spectral features

Spectral Centroid- Measures the spectrum's center of gravity. Calculated by dividing the average weighted frequency of amplitudes divided by sum of amplitudes. [20]

Spectral Entropy- Entropy of normalized spectral energies for a set of sub frames.

Spectral Flux- Squared difference of two normalized magnitudes of successive signal frames. [20]

Spectral Spread- Second central moment of spectrum

Spectral Rolloff- Frequency below which 90% of power of spectrum is concentrated.

2.2.4. Mel frequency Cepstral Coefficients (MFCC)

MFCC is one of the most commonly used feature set in Automatic Speech Recognition. It extracts a feature vector containing information about linguistic message. It mimics logarithmic perception of loudness and pitch of human auditory system, excludes fundamental frequencies and harmonics, hence eliminating speaker dependent characteristics. Current work makes use of 13 MFCC features.
The count of MFCC incorporates figuring the cosine transform of the real logarithm of the short span power range on a Mel scale.

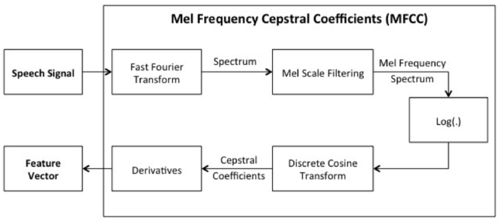

Figure 1: *MFCC feature extraction steps.*

Process of MFCC feature extraction is briefly listed [17]-

- Discrete Fourier transform (DFT) is applied on each frame, for converting it to frequency domain
- Next, Mel Frequency Cepstrum is calculated. Nd unique band pass filters filter the spectrum, and power of each frequency band is computed.
- Logarithm of signal is computed to mimic human perception of loudness
- Cepstral coefficients are computed, trying to eliminate speaker dependent characteristics
- First and Second order derivatives of cepstral coefficients extend the feature vector alongside the current frame

2.2.5. Chroma Features

These are related to 12 different pitch classes. They capture harmonic and melodic characteristics of music, and are robust to change in instrumentation. Current work makes use of 13 chroma features.
A human perceives two signals as of same pitch if they differ by an octave. Based on this, a pitch can be separated into two components, one of which is chroma. Twelve Chroma values are represented by the set

$$C, C, D, D, E, F, F, G, G, A, A, B$$

Set of all pitches that share the same chroma is called a Pitch Class. Because of this close relation, chroma features are also called 'Pitch Class Features'. [21]

2.2.6. Classifiers Used

Support Vector Machines (SVM) [c22] are supervised learning models with associated learning algorithms that dissect information and perceive designs, utilized for classification and regression. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.
k-Nearest Neighbor (kNN): k-nearest algorithm is based on the instance learning method. In this algorithm, it is assumed that instances belong to points in the multidimensional space.

The distance between neighbors are calculated using Euclidean distance[25].

Random Forests (RF) as machine learning algorithm has a major advantage- it can be used both as a classification and a regression problem. It is a supervised learning algorithm, and merges the predictions of multiple decision trees to reach a stable prediction. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model [24]

2.2.7. Egyptian Vulture Optimization

EVOA was a meta heuristic approach, initially developed to solve complex arrangement problems. It is aroused by the conduct of Egyptian vulture for procuring it's nourishment. The smart conduct of this winged creature is transformed into an algorithm which is capable of taking care of hard optimization problems. Egyptian Vulture is unique in relation to different winged animals in it's method for hunting. Their essential food is flesh yet as opposed to eating flesh from animals they eat flesh from the eggs of different birds. However to enter into big and hard eggs they need to utilize stones as sledge for breaking in. Another intriguing demonstration is moving with twigs. They move objects with twigs which is another particular highlight. These two primary activities of this remarkable bird takes the shape of this algorithm.

EVOA steps have been detailed and is adapted in the system

- Solution set or strings are initialized which have the representations of parameters and it is in the form of variables. A string that shows a set of parameters represents a single state of allowable solution.
- Representative variables are refined, superimposed constraints and conditions are checked.
- Stones are tossed at random or selected points.
- Selected part or entire string is chosen for Rolling of Twigs to be performed on.
- Selective part of solution is reversed using the tactic of change of angle
- Fitness is evaluated
- Stopping criteria should be checked

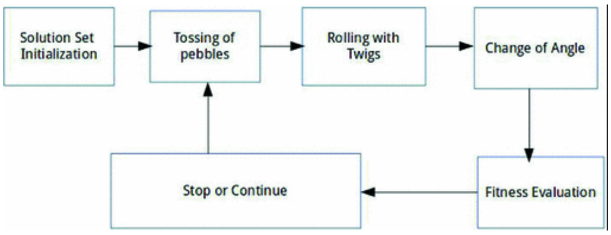

Figure 2: *EVOA Flow Chart*

2.2.8. Grey WOlf Optimization

GWO is a meta heuristic optimization and it is based on the pattern of the hunting followed by grey wolves [29]. There exists a social hierarchy in grey wolves which divides the group and help in hunting. There are four social group in grey wolves, alpha, beta, omega and delta. Alpha wolves are leaders, subordinates are beta and omega are lowest level wolves. Delta is a special category that consists of old wolves , trackers etc.

Grey wolf hunting involves tracking, chasing, and approaching prey slowly then they encircle the prey and harass it until prey stops movement and then attacks it. Following are the steps in the grey wolf optimization.

- Initialize set of parameters like grey wolf/ search agents numbers, design variable size etc.
- Generate randomly search agents(wolves) based on the size of the pack.
- Evaluate the fitness of grey wolves.
- Find best candidate agent based on the fitness value (alpha, beta and gamma)
- Move the agents or update the location according to the position of the alpha , beta and gamma.
- Repeat steps from 3 to 5 until stopping criteria met.

2.2.9. Moth Flame Optimization

Moths uses transverse orientation to fly in the night and this navigation depends on the light of the moon [26]. In transverse orientation a fixed angle is maintained with respect to the light source which in case of the moth is moon. By using this method, moths can travel in a straight line over a long distance this is because of the fact that moon is far away from the earth and the angle remain same [27]. But when this method is applied to the artificial light then it results in the spiral path as light source is very close and trying to make the angle results in the spiral type movement [28]. MFO is inspired from the spiral movement of the moth and tries to find the optimal solution in the local region of the flames(Best moth till now which acts as light source for other moths). Hence MFO search the space around the best solution obtained so far. Steps involved in the MFO are

- Initialize set of parameters like number of moths, shape of moths etc.
- Randomly generate moths according to the shape of moths
- Evaluate the fitness of the moths
- select best moths so far for the flames
- Update the location of the moth based on the flame it is following and spiral function. Location Update is given by
$$S(M_i, F_j) = D_i \cdot e^{bt} \cdot \cos(2\pi t) + F_j \qquad (1)$$
- Repeat steps from 3 to 5 till termination criteria not met.

3. METHOD IMPLANTED

Procedure and steps carried out in this work are described below:

1. Feature Extraction from dataset
   (a) Extract ZCR
   (b) Extract Energy Features
   (c) Extract Spectral Features
   (d) Extract MFCC Features
   (e) Extract Chroma Features
2. Extract statistical measure of features

3. Train upon Classifiers:
   (a) SVM
   (b) RF
   (c) KNN
4. Apply EVOA, GWO and MFO with above mentioned classifier accuracy as fitness function
5. Compare the results of EVOA, MFO, and GWO

4. Implementation and Results

This work tried to model Emotion Recognition from Speech using Nature Inspired Egyptian Vulture Optimization Algorithm.

4.1. Parameter settings

EVOA has been used with 170 features, 20 vultures optimizers, maximum 100 iterations and a switch probability of 0.8. GWO has been used with 170 features, 20 grey wolves optimizers, maximum 100 iterations. MFO has been used with 170 features, 20 moths optimizers, maximum 100 iterations . Parameters corresponding to three classifiers used are-

- **RF** - 100 estimators
- **KNN** - 5 nearest neighbors
- **SVM** - c = 100, kernel = RBF, decision function shape = one vs one

4.2. Experiment

All 170 features have been used. EVOA, GWO and MFO have been applied over three machine learning classifiers namely RF, KNN and SVM, and the comparative analysis of results of plain classifier vs EVOA vs GWO vs MFO optimized is shown below-

Table 1: *Accuracy: Comparison table*

Classifier (Cls.)	Only Cls.	Cls.+ EVOA	Cls.+ GWO	Cls.+ MFO
RF	71.22	83.33	81.48	81.35
KNN	70.8	87.03	85.18	80
SVM	77.82	88.88	90.74	86.66

From the experiments, it is seen that SVM with GWO produces the highest classification accuracy. In case of KNN and RF, combination with the GWO does not gave better result than EVOA. Fig 3 represents the accuracy with respect to iterations when GWO-SVM model applied. MFO algorithm performed worst among applied NIA algorithms.

5. CONCLUSIONS

This works attains the maximum testing accuracy of 90.70%, by applying Support Vector Machines with Grey Wolf Optimization over 7 emotional classes, taken from Berlin Emotional Corpus. This result is near comparable to previous works on the same database.

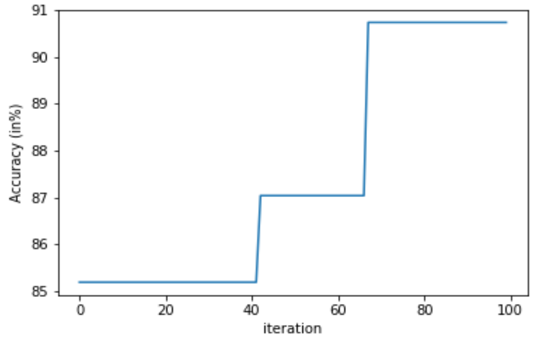

Figure 3: *Accuracy of the Best performing model with respect to Iterations. It shows the accuracy of the best search agent for the corresponding iteration.*

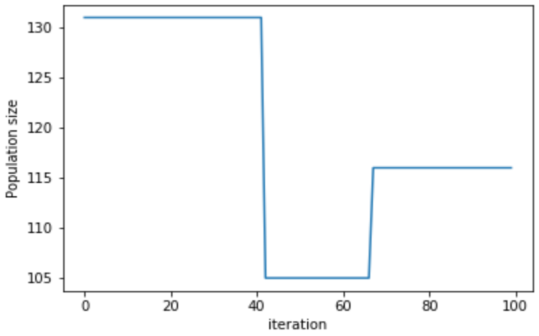

Figure 4: *Number of Features for the Best performing model with respect to Iterations. It shows the number of features in the best search agent for the corresponding iteration.*

6. References

[1] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. ”Survey on speech emotion recognition: Features, classification schemes, and databases.” Pattern Recognition 44.3 (2011): 572-587.

[2] Schuller, Bjrn, Gerhard Rigoll, and Manfred Lang. ”Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture.” Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on. Vol. 1. IEEE, 2004.

[3] France, Daniel Joseph, et al. ”Acoustical properties of speech as indicators of depression and suicidal risk.” IEEE transactions on Biomedical Engineering 47.7 (2000): 829-837.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expres- sions, IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 3958, Jan. 2009.

[5] I. Luengo, E. Navas, and I. Hernandez, Feature analysis and evalua- tion for automatic emotion identification in speech, IEEE Trans. Mul- timedia, vol. 12, no. 6, pp. 490501, Oct. 2010

[6] E. A. Moataz, K. M. S. , and K. Fakhri, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recog., vol. 44, no. 3, pp. 572587, 2011.

[7] Ekman, Paul. ”An argument for basic emotions.” Cognition & emotion 6.3-4 (1992): 169-200.

[8] Russell, James A., and Lisa Feldman Barrett. ”Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.” Journal of personality and social psychology 76.5 (1999): 805.

[9] Mao, Qirong, et al. ”Learning salient features for speech emotion recognition using convolutional neural networks.” IEEE Transactions on Multimedia 16.8 (2014): 2203-2213.

[10] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. ”Automatic speech emotion recognition using recurrent neural

networks with local attention." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[11] T. L. Nwe, S. W. Foo, and L. C. D. Silva, Speech emotion recognition using hidden Markov models, Speech Commun., vol. 41, no. 4, pp. 603623, 2003.

[12] Y. Sungrack and C. D. Yoo, Loss-scaled large-margin Gaussian mix- ture models for speech emotion classification, IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 2, pp. 585598, Feb. 2011.

[13] G. Davood, S. Mansour, N. Alireza, and G. Sahar, Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network, Neural Comput. Appl., vol. 21, no. 8, pp. 21152126, 2012.

[14] T. L. Pao, Y. T. Chen, J. H. Yeh, Y. M. Cheng, and Y. Y. Lin, A com- parative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech, Advanced Intell. Comput. Theories and Appl. With Aspects of Theoretical and Methodological, vol. 4681, pp. 9971005, 2007.

[15] Q. R. Mao and Y. Z. Zhan, Speech emotion recognition method based on improved decision tree and layered feature selection, Int. J. Hu- manoid Robot., vol. 7, no. 2, pp. 245261, 2010.

[16] Swain, Monorama, Aurobinda Routray, and P. Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." International Journal of Speech Technology (2018): 1-28.

[17] Becchetti C, Ricotti LP (2004) Speech recognition: theory and implementation, 3rd edn. Wiley, New York, pp 125135

[18] Theodoros Giannakopoulos(2017) "Python Audio Analysis Library: Feature Extraction, Classification, Segmentation and Applications" Online; accessed 5-March-2018

[19] Wikipedia contributors "Zero-crossing rate" Online; accessed 4-March-2018

[20] Swain, Monorama, Aurobinda Routray, and P. Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." International Journal of Speech Technology (2018): 1-28.

[21] Wikipedia contributors (2017) "Chroma feature — Wikipedia, The Free Encyclopedia" Online; accessed 5-March-2018

[22] Bo Yu, Haifeng Li and Chunying Fang, Speech emotion recognition based on op-timized support vector machine, Journal of software, VOL. 7, NO 12, December12., 2012.

[23] L. Fu, X. Mao, and L. Chen "Relative Speech Emotion Recognition Based Artificial Neural Network" Pacific-Asia Workshop on Computational Intelligence and Industrial Application, PACIIA, 2008.

[24] Niklas Donges "The Random Forest Algorithm" Online:accessed 14-May-2018 https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[25] Mitchell TM (1997) Machine learning. McGraw-Hill, Inc., New York

[26] S. Mirjalili, Moth-Flame Optimization Algorithm: A Novel Nature- inspired Heuristic Paradigm, Knowledge-Based Systems, Vol. 89, pp. 228-249, 2015

[27] K.D. Frank, C. Rich, T. Longcore, Effects of artificial night lighting on moths, Journal of Ecological consequences of artificial night lighting, Hindawi Publishing Corporation, Vol. 2015, No. 1, pp. 305-344, 2006

[28] K.J. Gaston, J. Bennie, T.W. Davies, J. Hopkins, The ecological im- pacts of nighttime light pollution: a mechanistic appraisal, Biological reviews, Hindawi Publishing Corporation, Vol. 88, No. 1, pp. 912-927, 2013

[29] Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Andrew Lewis. "Grey wolf optimizer." Advances in engineering software 69 (2014): 46-61