WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

Using Speech Signal for Emotion Recognition Using Hybrid Features with SVM Classifier

Fatima A. Hameed[1,*] and Loay E. George[2]

[1]Informatics Institute for Postgraduate Studies Iraqi Commission for Computer & Informatics, Iraq, Baghdad
[2]University of Information Technology and Communication

*Corresponding Author: Fatima A.Hameed

DOI: https://doi.org/10.31185/wjcm.102
Received: December 2022; Accepted: February 2023; Available online: March 2023

ABSTRACT: Emotion recognition is a hot topic that has received a lot of attention and study,owing to its significance in a variety of fields, including applications needing human-computer interaction (HCI). Extracting features related to the emotional state of speech remains one of the important research challenges. This study investigated the approach of the core idea behind feature extraction is the residual signal of the prediction procedure is the difference between the original and the prediction .hence the visibility of using sets of extracting features from speech single when the statistical of local features were used to achieve high detection accuracy for seven emotions. The proposed approach is based on the fact that local features can provide efficient representations suitable for pattern recognition. Publicly available speech datasets like the Berlin dataset are tested using a support vector machine (SVM) classifier. The hybrid features were trained separately. The results indicated that some features were terrible. Some were very encouraging, reaching 99.4%. In this article, the SVM classifier test results with the same tested hybrid features that published in a previous article will be presented, also a comparison between some related works and the proposed technique in speech emotion recognition techniques.

Keywords: Speech emotion recognition, Feature extraction, Statistical moments, Support vector machine (SVM)

CC BY

1. INTRODUCTION

The natural way to express ourselves as human beings is through speech. Human speech is the most significant information carrier for cognitive-communication, human identification, and emotional condition. Emotional intelligence has an essential role in human connection and communication. While everyone intuitively understands what emotions are, they are difficult to define. Emotional responses are critical in human-computer interaction. Recently, verbal emotion recognition has attracted increasing interest, which aims to analyze emotional states through speech signals [1]. There are six widely accepted archetypes of emotions based on psychology theory; anger, happiness, fear, sadness, and surprise. The human speaking tone and facial movements are essential in expressing feelings [2]. "Speech emotion recognition" is extracting a speaker's emotional situation from their speech. Speech emotion detection is thought to be beneficial for removing valuable semantics from speech and improving efficiency.

To investigate various aspects of speech signals. There are many applications in distinguishing verbal emotions when it is a joint work between man and machine, as they are helpful in the field of such intelligence help, as in criminal inquir [3], detection of frustration, disappointment, surprise/amusement [4], health care and medicine [5] and a better "Human Computer Interface" [6]. Also, it is beneficial for in-car board systems, where data about the driver's state of mind may be sent to the system to begin their protection. There are several essential kinds of research have been presented in this domain, and the primary difficulties encountered include choosing a speech database, finding distinct speech aspects,

and selecting the appropriate classification techniques.such as support vector machine (SVM) and artificial neural network (ANN) and different proposed approaches for speech emotion recognition [7].

S.Basu et al. (2017) [8] suggested a method for speech-based emotion recognition that employed the thirteen MFCCs and thirteen components of acceleration as features and a convolution neural network (CNN) with long short term memory (LSTM) for classification. About 80% of the accuracy was successfully achieved. When this method is fed a larger database, it can produce better results. M. S. Likitha et al. (2017) [9] employed MFCC features for feature extraction with SVM as a classification model and were successful in achieving the same accuracy.

P. Shegokar and P. Sircar (2016) [10] suggested a speech-based emotion recognition system in which the choice of features depends on the continuous wavelet transformation and prosodic coefficients Various SVMs are used as a classification model in the provided approach. The findings of the experiment indicate that 60.1% is the optimal rate of recognition.

A. Bhavan et al. (2019) [11] The extraction of a set of spectral characteristics (MFCCs and spectral centroids) that are preprocessed and reduced to the necessary set of features forms the basis of a suggested speech emotion identification technique. A bagged ensemble of SVMs with a Gaussian kernel was suggested as a classification model in the method that was just given. The accuracy ratio that was obtained is 84.11%. The language features (semantic features), which are not used in this technique because it is focused mainly on auditory features, may enhance its performance.

Z. Han and J. Wang (2017) [12] used SVM and Gaussian Kernel Nonlinear Proximal SVM to suggest a method for speech-based emotion identification. This method extracts the prosody and quality aspects of speech after preprocessing the voice signal. The final emotion recognition result is then obtained using a classification model that combines SVM and Proximal SVM. With SVM, the average recognition rate is 80.75%, and with Proximal SVM, it is 86.75%. These data demonstrate that the Proximal SVM approach has a higher rate of emotion recognition. Additionally, it is three times faster than SVM. To provide excellent outcomes, the suggested technique has to make use of more effective aspects.

Some related research aims to characterize emotional responses by studying vocal performance and features of an audio signal and data transmission. However, the speech-emotion recognition system needs to go through a few straightforward stages to be practical, quick, and accurate. Therefore, to reduce the system's complexity while maintaining high system accuracy, Using feature extraction methods and classifiers results in high fidelity and low complexity that simulates the human auditory system.

Feature extraction and classification is the most critical part of the system. In previous work [13] discussed a method of statistics for local features used in this study to achieve high detection accuracy.The proposed approach is based on the fact that local features can provide efficient representations suitable for pattern recognition. Hybrid features were evaluated in ANN, and competitive results were obtained. In this article, the same approach to extracting features from speech signals is discussed, using the same feature types proposed in previous works. The structure of this article is as follows: Section II describes the datasets that were used and the suggested techniques; Section III examines the findings of the experiments; Section VI reviews prior research that is relevant to this paper; and Section V offers conclusions.

2. THE PROPOSED METHODOLOGY

A crucial part of creating a system for speech emotion recognition (SER) is extracting characteristics that classify the best emotions. This study uses speech signals as a set of statistical features to build an emotion recognition system. In feature extraction, the statistics of local features were used to extract features by calculating the mean for a specific region. The basic concept is the residual signal of the prediction procedure, which is the difference between the original and its prediction. In a residual signal, the original signal is taken and subtracted from the local by extracting the mean, and the mean can be narrow or wide; this is the basic concept that has been worked out. This is done after applying the statistical, mathematical equations. To design and implement the local feature that can select the most appealing of features extracted, to achieve higher classification results in less time-consuming. Speech wave-based automatic emotion recognition and emotion state system apply four main stages (i.e., preprocessing, feature extraction, feature selection, and classification phase) to the input speech signal. The preprocessing includes step normalization. Feature extraction contains Spectral features and statistical moments. In feature selection, including selecting the most discriminative features. The SVM classifier is used for the classification task. The structure of the proposed system is shown with different proposed methods shown in Figure (1.1).

2.1 DATASETS

The publicly accessible Berlin dataset [accepted 2022/10/31 in Journal 'solid state phenomena' to (be published)], considered a popular Dataset of emotion recognition recommended in the literature, is public, and the quality of its recording is excellent. It contains 535 utterances spoken by ten actors (five males and five females) using ten texts. The

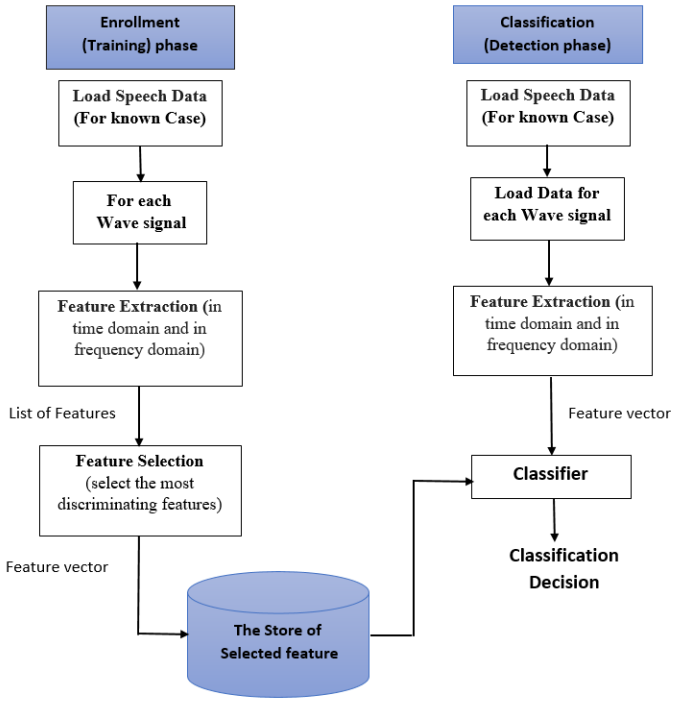

FIGURE 1. The architecture of the proposed system

Dataset is divided into Seven emotions "happy," "sadness," "Anger," "fear," "disgust," "Boredom," and "neutral" [14].

2.2 PROPOSED SYSTEM

The proposed methods in [13], which worked under the approach (using the Residual Signal of the prediction procedure, which is the difference between the original and its prediction), are tested in ANN in this study under the same approach (is tested in SVM).

2.2.1. Proposed Features

A set of Hybrid features were used by using the concept of residual signal for speech emotion recognition system in the previous study; these are the Prosodic Features like zero crossing, Daubechies Wavelet Transform (db4), and statistical moment features like(Moment ofActualWav,Moment of Residue ,Moment ofGradient,Moment ofAbsolute WavValues):

a) Prosodic Features

The rate at which a signal goes from positive to zero to negative (or vice versa) is referred to as zero crossing (ZCR), and it is a type of measurement feature." [15] [16] used in [13].

$$ZCR_{(n)} = \frac{1}{2(L-1)} \sum_{i=1}^{L-1} \left| \operatorname{sgn}(x_{i+1}) - \operatorname{sgn}(x_i) \right| \tag{1}$$

Where n is the number of input signals under processing, L is the signal length, and X_i is the i^{th} sample in each signal (n)

b) Daubechies Wavelet Transform (db4)

Daubechies wavelet also computes the sums and differences like HWT but differs from HWT in the scaling signals and wavelets. The values of scaling numbers that are used to obtain low coefficient are Daubechies Wavelet Transform (db4), which is described by (2) and (3) [17] [18] used in [13].

$$L(i) = \sum_{i=0}^{N/2} a(k)\, s(j+k) \tag{2}$$

$$H\left(i + \frac{N}{2}\right) = \sum_{i=0}^{N/2} \beta(k) s(k+j) \tag{3}$$

Where, i∈{0,.., (N/2)-1}, j∈{0,.., N-3}, and k∈{0,.., 3}. The scale values (α) and wavelets (β) are given below:

$$\alpha_1 = (1 + \sqrt{3}) /(4\sqrt{2}\ , \qquad \alpha_2 = (3 + \sqrt{3}) /(4\sqrt{2} \tag{4a}$$

$$\alpha_3 = (3 - \sqrt{3}) /(4\sqrt{2}\ , \qquad \alpha_4 = (1 - \sqrt{3}) /(4\sqrt{2} \tag{4b}$$

$$\begin{aligned} \beta_1 &= \alpha_4\ , \quad \beta_2 = -\alpha_3 \\ \beta_3 &= \alpha_2\ , \quad \beta_4 = -\alpha_1 \end{aligned} \tag{4c}$$

c) Statistical Moments Features

Moments measure the degree to which a particular quantity deviates considerably "from its mean or any pivot point in terms of mass, force, histogram intensity, frequency transform coefficients, or other types of coefficients with specific geometrical distributions". Mass, force, histogram intensity, frequency transform coefficients, and other types of coefficients can all be used to calculate moments. [19]. There are numerous different categories that moments can place under. Moment characteristics do calculate mathematically to characterize the object's behavior and extract key aspects. These features are described by [4–7] and [20–23]. All the features mentioned above have been used in the article [13]

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{6}$$

Where, N is the total number of samples S_i is the sample, μ is the mean.

$$p = \frac{1}{N}\sum_{i=1}^{N}(|x_i|)^2 \tag{7}$$

Where N is the total number of samples, x_i is the sample.

$$\mu' = \frac{1}{N}\sum_{i=0}^{N}(X_i - \bar{\mu})^o \tag{8}$$

Where o is the order of the moment (i.e., 1, 2, 3, 4), and $\bar{\mu}$ is the mean of Xi.

$$X_p = (\textstyle\sum_{i=1}^{n}(x_i|^p)\frac{1}{p} \tag{9}$$

Where p is the order of p-norm (i.e., 1/3, 1/6), and x_i is the mean.

2.2.2. Feature Selection

In this stage, various feature combinations have been examined, compared, and discussed in detail [13] to determine those that yield the highest achievable recognition rates.

3. CLASSIFICATION

Classification is the process of distinguishing class data from other sources of data in the feature space. There are different classifiers available for emotion recognition. In the statistical technique, patterns to be categorized are represented by a set of characteristics that define a multidimensional vector. In pattern classification, there are numerous approaches, such as the classic (statistical) approach, template matching, and syntactic [24].

Support Vector Machine (SVM) was first proposed by Vladimir Vapnik, along with Bernhard Boser and Isabelle Guyon, in 1992.SVM is a classification of the supervised aided methods because training needs certain learning objectives.

It is a commonly used classifier in a variety of pattern recognition challenges, such as emotion recognition and detection problems, and it is used to identify emotions in real-time for human speech [25].Also, this proposed methodology can also be used for object recognition, voice recognition, handwriting recognition, etc.

Support Vector Machine (SVM) is an extremely popular classification approach because of its accurate result and rapid computation. SVM is a method for classification and regression prediction [26]. SVM and ANN are comparable class-supervised learning techniques.

MATLAB is used to train an SVM algorithm for the classification of detections. In the present work, the cubic kernel is used with SVM. Following feature extraction, initially, data are divided into 80% and 20%. Testing data are taken as 20%. Out of the remaining 80%, training data are selected

4. TEST RESULTS

The results of a few tests conducted to evaluate the effectiveness of the established system are presented and discussed in this paper. MATLAB and the C# programming language of Microsoft Visual Studio 2017 were employed.

To test the accuracy of the proposed system with all proposed feature extraction methods Berlin datasets were used.The accuracy of the proposed system is tested using all of the feature extraction methods on the Berlin emotion datasets.Berlin datasets are relatively (7 classes and 535 utterances). The highest achieved system recognition rate was99.4% for some feature sets for each proposed feature extraction using all the datasets.

The training and testing results of the SVM algorithm will be displayed and compared with a some related works use SVM algorithm. Table (1) shows the training results of some features in the system with dataset samples tested in the SVM algorithm;

Table 1. The results of training and testing using all features (81 features) to classify seven classes

Features name	Feature number	All accuracy
All Features	81	99.4%
Features of StdDev	9	98.4%
Features of zero crossing	5	98.8%
Features of db4 wavelet	30	91.9%
Signal Power (power^2)	8	93.8%
Features of centralized moments (p-norm)	8	95.1%
Higher degree moments (power $^{2/3}$)	8	92.0%
Higher degree moments (power $^{1/3}$)	8	95.1%

Table 2. The results of training and testing sets using (db4 and ZRC) features

Features name	Feature number	All accuracy
Dd4 wavelet, zero crossing	35	99.8%

Table 3. The best-attained results of training and testing sets only one feature

Features name	Feature number	All accuracy
Standard Deviation	1	86.6%
Signal Power of ZRC	1	88.2%
p-norm of ZRC	1	87.5%
Power$^{2/3}$ of ZRC	1	91.1%
Power$^{1/3}$ of ZRC	1	94.1%
StdDev of Actual moments	1	90.2%
Power^2 of Residue	1	93.5%
Power of Gradient	1	89.0%
Power$^{2/3}$ of Absolute	1	95.4%
Power$^{1/3}$ of Gradient	1	94.3%

5. COMPARISON WITH RELATED WORKS

Although many of the published research on speech emotion recognition systems used more than one feature to identify the emotional state, several of them showed encouraging results. Table (8) compares the suggested technique and another

Table 4. The best-attained results of training and testing sets only two features

Features name	Feature number	All accuracy
Power of ZRC, $Power^{1/3}$of ZRC	2	89.5%
$Power^2$ of ZRC ,$Power^{2/3}$ of ZRC	2	95.1%

Table 5. The best-attained results of training and testing sets only three features

Features name	Feature number	All accuracy
StdDev of ZRC, $Power^2$ of ZRC, $Power^{2/3}$of ZRC	3	67.8%
StdDev of ZRC, Power of ZRC, $Power^{1/3}$ of ZRC	3	94.3%
StdDev of Gradient ,$Power^{2/3}$ of Actual , $Power^{1/3}$ of Residue	3	94.9%

Table 6. The best-attained results of training and testing sets only four features

Features name	Feature number	All accuracy
$Power^2$, $Power^{2/3}$, P-norm, $Power^{1/3}$ of ZRC	4	89.2%
StdDev of Gradient, StdDev of ZRC, $Power^{1/3}$ of Residue, $Power^{2/3}$ of Absolute	4	96.2%
$Power^2$of Gradient , $Power^{2/3}$ of Actual $Power^{1/3}$ of ZRC,power of db4	4	79.0%

Table 7. The best-attained results of training and testing sets only five features

Features name	Feature number	All accuracy
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet0	5	89.7%
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet1	5	82.9%
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet2	5	95.5%
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet3	5	92.0%
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet4	5	95.6%
StdDev, $power^2$, $power^{2/3}$, $power^{1/3}$, p-norm of wavelet5	5	96.2%

few related works in speech emotion recognition. This table illustrates how the suggested technique enhanced accuracy more than the majority of alternative techniques.

Table 8. A comparison between the proposed technique and some related works of speech emotion recognitiontechniques.

Author/(s), Year, Reference	**The used Features**	**The used Classifier**	**Accuracy**
P. Shegokar and P. Sircar 2016, [10]	Transformation of continuous wavelet and Coefficients of prosodic	SVMs	60.1%
Basu et al. 2017, [8]	MFCC	CNN with LSTM	80%
M. S. Likitha et al. 2017, [9]	MFCC	SVM	80%
Z. Han and J. Wang 2017, [12]	Features of speech prosody and quality	SVM, Proximal SVM	80.75%, 86.75%
A. Bhavan et al. 2019, [11]	MFCCs and spectral centroids	SVMs	84.11%
Proposed Work	Dd4 wavelet, ZRC, std, Po^2, $Po^{2/3}$, po, $po^{1/3}$	SVM	99.4%

6. CONCLUSIONS AND FUTURE WORK

This article adopts a method for extracting features from user speech signals; the features proposed in previous studies are tested to check the degree of these traits' discrimination when tested in the SVM algorithm. This approach has excellent results in the emotion recognition system, but the performance of the SVM algorithm for proposed work by using Hybrid Features is better than some related works. Also, this strategy employs hybrid features and keeps the computational complexity low. This research showed that the statistics of local features used to extract features by computing the mean for a specific location are sufficient to extract distinguish features and recognize emotion when the suggested method was evaluated on accessible datasets.A novel statistical momentis recommended as a new feature for the speech emotion recognition system and can be tested on other data sets.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit*, vol. 44, no. 3, pp. 572–587, 2011.
[2] I. Chiriacescu, "Automatic Emotion Analysis Based on Speech," 2010.
[3] K. Wang *Speech Emotion Recognition Using Fourier Parameters*, vol. 6, pp. 69–75, 2015.
[4] D. D. Joshi and M. B. Zalte *Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms*, pp. 59–63, 2013.
[5] R. Subhashree and G. N. Rathna, "Speech emotion recognition: Performance analysis based on fused algorithms and GMM modeling," *Indian J. Sci. Technol*, vol. 9, no. 11, 2016.
[6] A. Milton, S. Roy, and S. Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," *Int. J. Comput. Appl*, 2013.
[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.
[8] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using a convolutional neural network with recurrent neural network architecture," *2017 2nd International Conference on Communication and Electronics Systems (ICCES*, pp. 333–336, 2017.
[9] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech-based human emotion recognition using MFCC," 2018.
[10] P. Shegokar and P. Sircar, "Continuous wavelet transform-based speech emotion recognition," 2016.
[11] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Syst*, vol. 184, pp. 104886–104886, 2019.
[12] Z. Han and J. Wang, "Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine," 2017.
[13] F. A. Hameed and L. E. Georgeb, "Using Speech Signal for Emotion Recognition Using Hybrid Features with ANN Classifier, accepted 2022/10/31 in journal 'solid state phenomena' to," 2022.
[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," 2005.
[15] R. W. Wall, "Simple methods for detecting zero crossing," *IECON'03. 29th Annual Conference of the IEEE Industrial Electronics Society*, vol. 3, pp. 2477–2481, 2003.
[16] F. Alías, J. C. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Appl. Sci*, vol. 6, no. 5, pp. 143–143, 2016.
[17] C. Vonesch, T. Blu, and M. Unser, "Generalized Daubechies wavelet families," *IEEE Trans. Signal Process*, vol. 55, no. 9, pp. 4415–4429, 2007.
[18] J. S. Walker, "A primer on wavelets and their scientific applications," 2008. Chapman and hall/CRC.
[19] D. Dacunha-Castelle and M. Duflo *Probability and Statistics*, vol. II, 2012.
[20] A. B. Downey, *Think Stats Probability and Statistics for Programmers. Version 1.6. 0.* Massachusetts: Green Tea Press, 2011.
[21] G. Bohm and G. Zech, *Introduction to statistics and data analysis for physicists*, vol. 1. Hamburg: Desy, 2010.
[22] H. A. Hadi, L. E. George, " Eeg, User, Methods, Two, Sets, Features, On, and Wavelet *J. Theor. Appl. Inf. Technol*, vol. 95, no. 22, 2017.
[23] R. W. Grubbström and O. Tang, "The moments and central moments of a compound distribution," *Eur. J. Oper. Res*, vol. 170, no. 1, pp. 106–119, 2006.
[24] C. G. V. N. Prasad, K. H. Rao, D. Pratima, and B. N. Alekhya, "Unsupervised Learning Algorithms to Identify the Dense Cluster in Large Datasets," *Int. J. Comput. Sci. Telecommun*, vol. 2, no. 4, pp. 26–31, 2011.
[25] M. H. Abdul-Hadi and J. Waleed, "Human speech and facial emotion recognition technique using SVM," *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 191–196, 2020.
[26] I. E. Naqa and M. J. Murphy, "What are machine learning? " in machine learning in radiation oncology, 3-11," 2015. Springer.