Emotion Recognition Based on Speech Signal Using Hybrid Features with Decision tree

Fatima A. Hameed
Informatics Institute for Postgraduate Studies/Iraqi Commission for Computers & Informatics (IIPS/ICCI)
Ms202010636@iips.icci.edu.iq
Dr. Loay E. George
University of Information Technology and Communication
loayedwar57@uoitc.edu.iq

Abstract :

Emotion recognition based on a speech signal is one of intensively studied research topics in the domains of human-computer interaction and affective computing. The main idea are a new Hybrid feature set was introduced in extract features , which use the basic concept in work is the Residual Signal of the prediction procedure, which is the difference between the original and its prediction ,by enter to the internal structure of the signal axis and calculate a features for some specific regions after that calculate moments for these regions . Machine learning techniques will be used to classify to achieve a high degree of accuracy.Publicly available speech datasets like the berlin dataset are tested using a decision tree classifier. The hybrid features were trained separately. The results indicated that features very encouraging, reaching 98.5%. In this article, the decision tree 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, Decision tree.

1.Introduction

The natural way to express ourselves as human beings is through speech.For cognitive communication, human identification, and emotional state, human voice is the most significant information carrier. The ability to connect and communicate with others requires emotional intelligence. Although while everyone has a basic understanding of what emotions are, they are challenging to define. In human-computer interaction, emotional reactions are crucial.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.

Feature extraction and classification is the most critical part of the system. In previous work [21] 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.Related Work

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 Decision tree and artificial neural network (ANN) and different proposed approaches for speech emotion recognition [3].

Yüncü, et al (2014)[4] The previous generation proposes to categorize seven emotions. The output of the computational model of the auditory system serves as the foundation from which features are extracted. The average and standard deviation of the output signals from the auditory model are the retrieved features. Binary decision tree classifiers are used to perform the classification. The outcomes showed that the suggested algorithm performs effectively for speaker independent situations and for various languages. Berlin Emotional Speech Database leave one sample out cross validation, which has an automatic recognition rate of 82.9%, has the highest accuracy. Polish Emotional Speech Database speaker independent results have the lowest accuracy, at 56.25%.

Liu,et al. (2018)[5] According to how confused different basic emotions are with one another, an extreme learning machine (ELM) decision tree-based solution for emotion recognition is suggested. A framework for speech emotion recognition is put forth, and classification tests are conducted utilizing the Mandarin speech database from the Institute of Automation of the Chinese Academy of Sciences (CASIA). The trial findings also indicate that the proposal had an average recognition rate of 89.6%. The proposed method would make it quick and easy to identify different speakers' emotional states from their speech.

Linhui, et al.(2019)[6] offer a good approach to speech emotion identification based on DNN-decision trees SVM from two perspectives: how to identify more different speech emotion features and build a strong recognition model. In the suggested method, speech emotion recognition is included. By determining the level of emotional confusion, a decision tree SVM framework is first constructed. In order to obtain bottleneck features that were utilized to train each SVM classifier in the decision tree, various DNNs were trained for various emotion groups. The results of the studies indicate that, when compared to traditional SVM and DNN-SVM classification methods, the average recognition rate of the proposed technique based on DNN-decision tree SVM may reach 75.83% higher.

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.

3.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 decision tree classifier is used for the classification task. The structure of the proposed system is shown with different proposed methods shown in Figure (1.1).

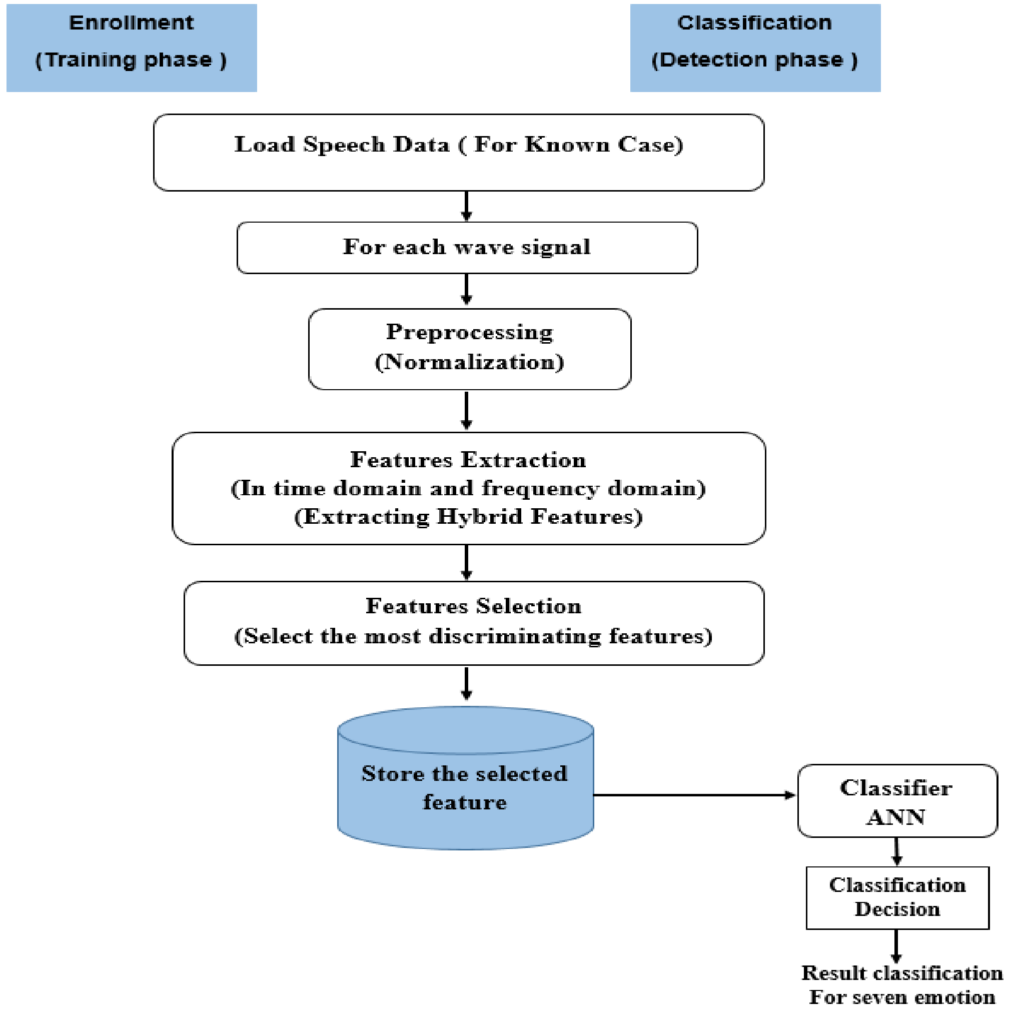

Figure (1.1): The architecture of the proposed system

3.1 Datasets

The publicly accessible Berlin dataset [21], 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 Dataset is divided into Seven emotions "happy," "sadness," "Anger," "fear," "disgust," "Boredom," and "neutral"[7].

3.2 The Proposed System

The proposed methods in [21], 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 decision tree).

3.2.1 Features Extraction

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) Zero crossing (ZCR)

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. used in [21]. ZCR can be defined as in equation (1) [8][9]:

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

Where n is the number of input signals under processing, L is the signal length, and X_i is the ith 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) used in [21], which is described by (2) and (3) as in equation [10][11] :

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

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

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

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

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

$$\beta_1 = \alpha_4\, , \quad \beta_2 = -\alpha_3 \qquad (4c)$$

$$\beta_3 = \alpha_2\, , \quad \beta_4 = -\alpha_1 \qquad (4d)$$

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. [12]. 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 (5,6,7,8,9)in [13][14][15][16][17]. All the features mentioned above have been used in the article [21].

$$\sigma = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(x_i - \mu)^2} \qquad (5)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (6)$$

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

$$p = \frac{1}{N}\sum\nolimits_{i=1}^{N}(|x_i|)^2 \qquad (7)$$

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

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

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

$$X_p = (\sum_{i=1}^{n}|x_i|^p)^{1/p} \qquad (9)$$

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

3.2.2 Feature Selection

In this stage, various feature combinations have been examined, compared, and discussed in detail [21] to determine those that yield the highest achievable recognition rates.

3.3.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[18].

One of the most well-known approaches to representing classifiers for data classification is the decision tree approach. The improved view of performance outcomes offered by decision tree classifiers is well established. All reputable data classifiers use increased tree pruning approaches, optimized splitting parameters, and strong precision[19].

Decision tree (DT)are robust, successful, and also very efficient tools used in a variety of fields, such as data mining, machine learning, information extraction, text mining, and pattern recognition. as well as continually finding methods to make the process more cost-effective, easier, accurate and efficient[20].

MATLAB is used to train an decision tree algorithm for the classification of detections. In the present work, the medium tree kernel is used with decision tree. 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 was 98.5% for some feature sets for each proposed feature extraction using all the datasets.

The training and testing results of the decision tree algorithm will be displayed and compared with a some related works use decision tree algorithm. Table (1) shows the training results of some features in the system with dataset samples tested in the decision tree 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	98.5%
Features of StdDev	9	94.4%
Features of zero crossing	5	96.8%

Features of db4 wavelet	30	97.0%
Signal Power (power^{2})	8	93.8%
Features of centralized moments (p-norm)	8	89.1%
Higher degree moments (power $^{2/3}$)	8	92.0%
Higher degree moments (power $^{1/3}$)	8	93.1%

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	97.5%

Table (3): The best-attained results of training and testing sets only one feature

Features name	Feature number	All accuracy
Standard Deviation	1	88.2%
Signal Power of ZRC	1	89.0%
p-norm of ZRC	1	86.5%
Power$^{2/3}$ of ZRC	1	91.1%
Power$^{1/3}$ of ZRC	1	92.3%
StdDev of Actual moments	1	90.2%
Power^{2} of Residue	1	78.0%
Power of Gradient	1	81.0%
Power$^{2/3}$ of Absolute	1	93.4%
Power$^{1/3}$ of Gradient	1	91.3%

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	85.5%
Power^{2} of ZRC ,Power$^{2/3}$ of ZRC	2	90.1%

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	87.8%
StdDev of ZRC, Power of ZRC, Power$^{1/3}$ of ZRC	3	90.3%
StdDev of Gradient ,Power$^{2/3}$ of Actual , Power$^{1/3}$ of Residue	3	91.9%

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	90.2%
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	92.0%

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	89.9%
StdDev, power^{2}, power$^{2/3}$, power$^{1/3}$, p-norm of wavelet1	5	88.9%
StdDev, power^{2}, power$^{2/3}$, power$^{1/3}$, p-norm of wavelet2	5	90.5%
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	95.3%
StdDev, power^{2}, power$^{2/3}$, power$^{1/3}$, p-norm of wavelet5	5	97.2%

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 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 some of related works and the proposed technique of speech emotion recognition techniques.

Author/(s), Year, Reference	The used Classifier	Accuracy
Yüncü, et al (2014)[4]	Binary decision tree	82.9%
Liu,et al. (2018)[5]	extreme learning machine (ELM) decision tree	89.6%
Linhui, et al.(2019)[6]	DNN-decision trees	85. 83%
Proposed Work	Decision tree	98.5%

6. Conclusion 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 decision tree algorithm. This approach has excellent results in the emotion recognition system, but the performance of the decision tree 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.

References

[1] M. El 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,” *Delft University*, 2010. .

[3] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116. Elsevier B.V., pp. 56–76, Jan. 01, 2020, doi: 10.1016/j.specom.2019.12.001.

[4] E. Yüncü, H. Hacihabiboglu, and C. Bozsahin, “Automatic speech emotion recognition using auditory models with binary decision tree and svm,” in *2014 22nd international conference on pattern recognition*, 2014, pp. 773–778.

[5] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree,” *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[6] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, “Speech emotion recognition based on DNN-decision tree SVM model,” *Speech Commun.*, vol. 115, pp. 29–37, 2019.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” 2005, doi: 10.21437/interspeech.2005-446.

[8] R. W. Wall, “Simple methods for detecting zero crossing,” in *IECON’03. 29th Annual Conference of the IEEE Industrial Electronics Society (IEEE Cat. No. 03CH37468)*, 2003, vol. 3, pp. 2477–2481.

[9] 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, p. 143, 2016.

[10] C. Vonesch, T. Blu, and M. Unser, “Generalized Daubechies wavelet families,” *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4415–4429, 2007.

[11] J. S. Walker, *A primer on wavelets and their scientific applications*. Chapman and hall/CRC, 2008.

[12] D. Dacunha-Castelle and M. Duflo, *Probability and Statistics: Volume II*, vol. 2. Springer Science & Business Media, 2012.

[13] A. B. Downey, “Think Stats Probability and Statistics for Programmers. Version 1.6. 0.” Massachusetts, Green Tea Press, 2011.

[14] G. Bohm and G. Zech, *Introduction to statistics and data analysis for physicists*, vol. 1. Desy Hamburg, 2010.

[15] J. Cohen, “Statistical power analysis,” *Curr. Dir. Psychol. Sci.*, vol. 1, no. 3, pp. 98–101, 1992.

[16] 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.

[17] H. A. Hadi and L. E. George, “EEG Based User Identification and Verification Using the Energy of Sliced DFT Spectra,” *Int. J. Sci. Res.*, vol. 6, no. 9, pp. 46–51, 2017.

[18] 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.

[19] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[20] Priyanka and D. Kumar, “Decision tree classifier: a detailed survey,” *Int. J. Inf. Decis. Sci.*, vol. 12, no. 3, pp. 246–269, 2020.

[21] Hammed,F.A.,&Georgeb,L.E.(2022).Using Speech Signal for Emotion Recognition Using Hybrid Features with ANN Classifier , acccepted 2022/10/31 in journal 'solid state phenomena ' to (be publish