International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 21st Jul 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 06

10.48047/IJIEMR/V12/ISSUE 06/03

Title **HUMAN EMOTION RECOGNITION BASED ON SPEECH SIGNAL**

Volume 12, ISSUE 06, Pages: 13-18

Paper Authors **J. Pradeep, K. Shivakrishna, N.Sushanth, Dr. B. Balnarsaiah**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

HUMAN EMOTION RECOGNITION BASED ON SPEECH SIGNAL

J. Pradeep, K. Shivakrishna, N.Sushanth guided by **Dr. B. Balnarsaiah**,
Assistant professor, Department of ECE, Nalla malla reddy engineering college, Telangana, India.
jittaboinapradeep123@gmail.com, shivakrishnachary5@gmail.com, Sushanthreddynalabolu@gmail.com

***ABSTRACT*:**The paper presents speech emotion recognition from speech signals based on features analysis and NN-classifier. Automatic Speech Emotion Recognition (SER) is importantfor measuring people's emotions in HCI systems. Ithas dominated psychology by linking expressions to basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise). The recognition system involves speech emotion detection, features extraction and selection, and classification. These features help distinguish the maximum number of samples accurately, and the PNN classifier based on discriminate analysis is used to classify the six different expressions. The simulated results will show that the filter-based feature extraction with the used classifier gives much better accuracy with lesser algorithmic complexity than other speech emotion expression recognition approaches.
Keywords *– MLP-Classifier, MFCC, Model, Neural Networks, Prediction.*

1. INTRODUCTION

SER is one of the booming research topics in the computer science world. Emotion is a medium by which one expresses how a person feels and one's state of mind. Emotions play an important factor in sensitive job areas, like that of a surgeon, a Military Commander, and many others where one must maintain their emotions. Predicting emotions is a tough task, as every individual has a different tone and intonation of speech. They elicit different emotions: happiness, anger, neutrality, sadness, and surprise. This paper aims to classify these emotions from a given speech sample in the most appropriate method. With different methods of predicting emotions, we plan to use the multilayer perceptron. We compare the two classifiers, i.e., Support Vector Machine (SVM) and Multilayer Perceptron Classifier

(MLP Classifier). SVM efficiently predicts emotions for sound input with no discrepancy. In the presence of noisy input, it deviates from its prediction. SVM only classifies using a single plane and restricts the prediction. The results show that the SVM system, i.e.,it has more computational time, even with decent accuracy. SVM only works on a single plane, so it faces problems addressing complex time series-based data.

Deep Neural Networks (DNNs) are now state-of-the-art machine learning models across various areas, from image analysis to Natural Language Processing, and are widely deployed in academia and industry. These developments have a huge potential for medical imaging technology, medical data analysis, medical diagnostics, and healthcare in general slowly being realized. We briefly overview recent advances and associated challenges in machine learning applied to medical image processing and image analysis. Long before deep learning was used, traditional machine learning methods were mainly used. Such as Decision Trees, SVM, Naïve Bayes Classifier and Logistic Regression. These algorithms are also called flat algorithms. Flat here means these algorithms cannot normally be applied directly to the raw data (such as .csv, images, text, etc.). We need a preprocessing step called Feature Extraction. The Feature Extraction result represents the raw data that these classic Machine Learning Algorithms (MLA) can now use to perform a task. For example, the classification of the data into several categories or classes. Feature extraction is usually quite complex and requires detailed problem-domain knowledge. This preprocessing layer must be adapted, tested, and refined over several iterations for optimal results. On the other side are the Artificial Neural Networks (ANNs) of Deep Learning. These do not need the Feature Extraction step. The layers can learn an implicit representation of the raw data directly and independently. Here, a more and more abstract and

compressed representation of the raw data is produced over several layers of artificial neuralnets. This compressed input data representation is then used to produce the result. The result can be, for example, the classification of the input data into different classes.

2. LITERATURE REVIEW

GMM Supervector Based SVM With Spectral Features For Speech Emotion Recognition:

Speech emotion recognition is a challenging yet important speech technology. This paper applies the Gaussian Mixture Model (GMM) super vector-based SVM with spectral features to this field. A GMM is trained for each emotional utterance, and the corresponding GMM super vector is used as the input feature for SVM. Experimental results on an emotional speech database demonstrate that the GMM super vector-based SVM outperforms standard GMM on speech emotion recognition.We cannot find the confused emotional states or other features, such as prosodic and voice quality features, in this method.

Fuzzy Rule-Based Voice Emotion Control for User Demand Speech Generation of Emotion Robot:

The emotional function of the human mind has an important role in decision-making, memory, action, and good communication. Especially emotionalvoice characteristics are very important for warm communication, successful business, good human-to-human relationship, and good care for children and silver ages. On the other hand, the service robot market such as educator, helper, secretary, deliver, and guider has been growing because of the old population and complicated social situation. In that case, the emotion function is needed in those areas. The emotional characteristic of voice depends on pitch contour, acoustic energy, vocal tract features, and speech energy. Therefore, we must consider how we must apply and implement the emotional function of voice for service robots. However, its implementation for robots is very difficult, and recognition is also not easy because of various emotional patterns in voice. This article suggests the method of voice emotion generation for user demand emotion talk in service robots. A fuzzy rule-based approach is introduced to generate emotion for user demand emotional function by controlling pitch contour, acoustic energy, vocal tract features, and speech energy.Voice emotion recognition is important for emotion robots and has many useful applications in other areas. In a real robot society, robots can be taught to interact with humans and recognize human emotions for communication. As the service robot grows, robotic pets, delivery robots, and care robots, for example, should be able to understand emotional situations, not only spoken commands.

Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition:

We propose integrating the attention mechanism into Deep Recurrent Neural Network (DRNN) models for speech emotion recognition. This is based on the intuition that it is beneficial to emphasize the expressive part of the speech signal for emotion recognition. By introducing an attention mechanism, the system learns how to focus on the more robust or informative segments in the input signal. The proposed recognition model is evaluated on the FAU-Aibo tasks defined in Interspeech 2009 Emotion Challenge.

The experiment results, the effect of using the attention mechanism is remarkable. Performance measured by UA recall rate improves from 37.0% of the RNN model to 46.3% of the LSTM-attention model. The main reason for the difference is the ability to locate and focus on the salient or reliable parts of the signal. From the distribution of the attention weights, we can see that the middle part of an utterance is often more important than the beginning/ending parts

An Investigation of Emotion Changes from Speech:

Emotion recognition based on speech plays an essential role in Human-Computer Interaction (HCI), which has motivated extensive recent investigation into this area. However, current research on emotion recognition is focused on recognizing emotion on a per-file basis and mostly does not provide insight into emotion changes. My research will investigate the emotion transition problem, including localizing emotion change points, recognizing emotion transition patterns, and predicting or recognizing emotion changes. As well as being potentially important in applications, the research delving into emotion changes paves the way towards a better

understanding of emotions from engineering and potentially psychological perspectives.

Our results show that emotion change points can be effectively detected when the proposed methods are combined, giving an EER as low as 20.8%. Currently, in this moving toward the next phase of my research, which is emotion change modeling, trialing on the IEMOCAP and the SEMAINE databases.

Construction of a Database of Emotional Speech Using Emotional Sounds from movies and Dramas:

In this study, an emotional speech database called Hanbat Emotional Database (HEMO) was constructed using movie and drama scenes in which professional actors abundantly express emotion. HEMO consists of 454 speech samples classified into seven emotion categories: anger, happiness, sadness, disgust, surprise, fear, and neutral. To evaluate the performance of HEMO, consistent experiments were conducted based on the Hidden Markov Model (HMM) and GMM for both HEMO and the Berlin Emotional Speech Database (EMO). HEMO showed better results than EMO, with a positive recognition rate of 78.89%.

A database of emotional speech was constructed using scenes in which professional actors produce abundantly expressed emotion in movies and dramas. The constructed emotional speech database, called HEMO, consists of 454 speech samples classified into seven emotion categories anger, happiness, sadness, disgust, surprise, fear, and neutral.

3.METHODOLOGY

Existing method

1. Principal Component Analysis (PCA)
2. Geometric Methods (GM)
3. SVM Classification

Principal Component analysis

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into values of linearly uncorrelated variables called **principal components**. The number of principal components is less than or equal to the number of original variables. This transformation is defined so that the first principal component has the largest possible variance (i.e., it accounts for as much of the variability in the data as possible). Each succeeding component, in turn, has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also named the discrete Karhunen-Loève transform (KLT) (ore Hotelling Transform), or Proper Orthogonal Decomposition (POD).

SVM classification

In ML, SVMs, are learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model represents the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that space and predicted to belong to a category based on which side of the gap they fall on.

A SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since the larger the margin, the lower the generalizationerror of the classifier.

Whereas the original problem may be stated in a finite-dimensional space, the sets to discriminate are often not linearlyseparable in that space. For this reason, the original finite-dimensional space was proposed to be mapped into a much higher-dimensional space, presumably making the separation easier in that space. The mappings employed by SVM schemes are created to make sure that dot products may be computed simply in terms of the variables in the original space by defining them in terms of a kernel function $K(x,y)$. This is done in order to keep the computational burden manageable.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

selected to suit the problem. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

Drawbacks

Low discriminatory power and high computational load

In geometric-based methods, geometric features like the distance between speech signals

Proposed method

Through textural analysis and NN classifiers, speech emotion recognition for transform features system is accomplished.

Advantages

Robustness to illumination changes

Low complexity

High discriminatory power

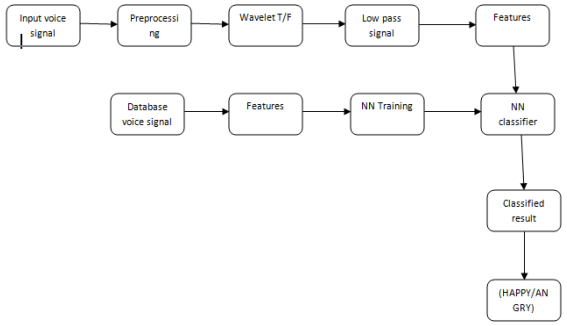

Fig.1: System architecture

MODULES:

Preprocessing

Operations involving signals at the lowest level of abstraction are frequently referred to as "preprocessing."

preprocessing aims to improve the signal data that suppresses unwanted distortions or enhances some signal features important for further processing.

The high-pass pre-emphasis filter can then equalize the energy between speech's low and high-frequency components.

Here, we are processing the person's speech signal. This process is based on the frequency level of the signal. Every person's audio signal,which means the speech varies by their tone of voice, and some expressions also differ. These are processed further before we process the input audio in the preprocessing unit.

DWT

The Discrete Wavelet Transform (DWT) provides a sparse representation of many natural signals. In other words, the important features of many natural signals are captured by a subset of DWT coefficients that is typically much smaller than the original signal. This "compresses" the signal. With the DWT, you always end up with the same number of coefficients as the original signal, but many of the coefficients may be close to zero. As a result, you can often throw away those coefficients and still maintain a high-quality signal approximation.

The strict discretization of scale and translation in the DWT ensures that the DWT is an orthonormal transform (when using an orthogonal wavelet). There are many benefits of orthonormal transforms in signal analysis, and many signal models consist of some deterministic signal plus white Gaussian noise.

Feature extraction

Feature extraction is a type of dimensionality reduction that efficiently represents interesting parts of an image as a compact feature vector. This approach is useful when large image sizes and a reduced feature representation are required to complete tasks such as image matching and retrieval quickly. This research evaluated speaker verification performance based on different MFCC feature extraction methods.

MFCC FEATURES:

The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method is a leading approach for speech feature extraction, and current research aims to identify performance enhancements.

The extracted MFCC feature vectors did not accurately capture the transitional characteristics of the speech signal, which contains the speaker-specific information. • Improvements in the characteristic transitional capture were found by computing DMFCC and DDMFCC, which were obtained respectively from the first-order and second-order time-derivative of the MFCC.

NEURAL NETWORKS:

Neural Networks (NN) and General Regression Neural Networks (GRNN) have similar architectures. Still, there is a fundamental difference: networks

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

perform classification where the target variable is categorical, whereas general regression neural networks perform regression where the target variable is continuous. If you select a NN/GRNN network, DTREG will automatically select the correct type of network based on the type of target variable.

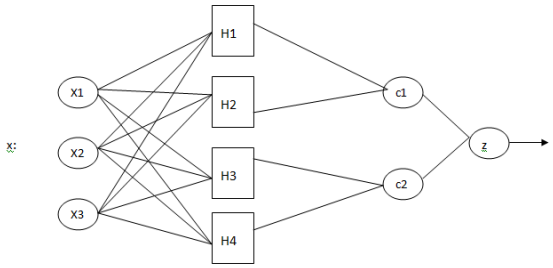

Fig.2: NN architecture

All NN networks have four layers:

1. Input layer — Each predictor variable has one neuron in the input layer. In the case of categorical variables, N-1 neurons are used where N is the number of categories. The input neurons (or processing before the input layer) standardize the range of the values by subtracting the median and dividing by the interquartile range. The input neurons then feed the values to each of the neurons in the hidden layer.

2. Hidden layer — This layer has one neuron for each case in the training data set. The neuron stores the values of the predictor variables for the case along with the target value. When presented with the x vector of input values from the input layer, a hidden neuron computes the Euclidean Distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma value(s). The resulting value is passed to the neurons in the pattern layer.

3. Pattern layer / Summation layer — The next layer in the network is different for NN networks and GRNN networks. For NN networks, there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent (hence, it is a weighted vote for that category). For GRNN networks, there are only two neurons in the pattern layer. The denominator summation unit is one neuron, while the numerator summation unit is another neuron.The denominator summation unit adds up the weight values of each hidden neuron. The numerator summation unit adds the weight values multiplied by the actual target value for each hidden neuron.

4. Decision layer — The decision layer differs for NN and GRNN networks. For NN networks, the decision layer compares the weighted votes for each target category accumulated in the pattern layer and uses the largest vote to predict the target category.

4. EXPERIMENTAL RESULTS

The below picture shown below refers to the filtered signal of input signal

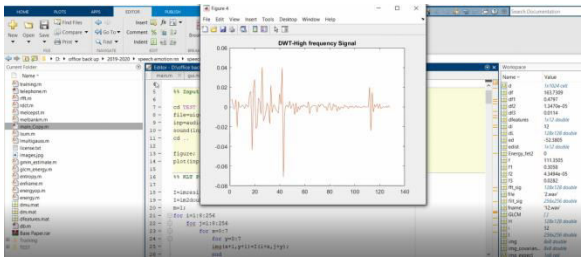

Fig.3: Output

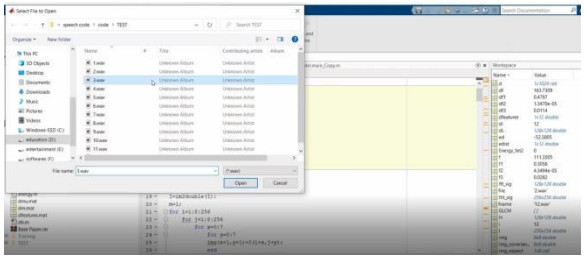

Fig.4: Output

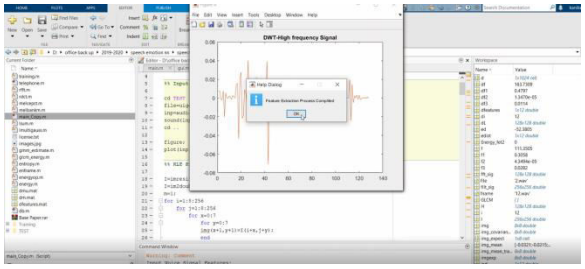

Fig.5: Output

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

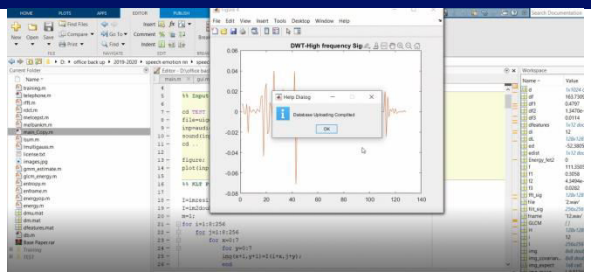

Fig.6: Output

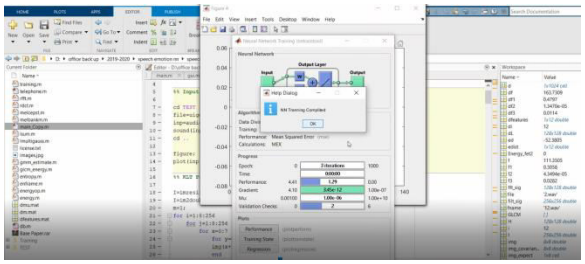

Fig.7: Output

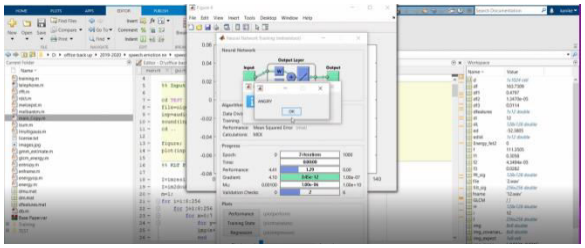

Fig.8: The above picture shows the emotion detected

6. CONCLUSION

• The voice emotion recognition from speech signals presented in this study is based on features analysis and NN-classifier. The simulated results will demonstrate that, compared to alternative spoken emotion expression identification systems, the filter-based feature extraction with the used classifier provides significantly improved accuracy with a more straightforward algorithm.

7. FUTURE SCOPE

In the future we increased the performance of this process and got more accuracy.

REFERENCES

[1] T. U. Binbin, "Speech emotion recognition based on improved rnfcc with emd," Computer Engineering and Applications, vol. 48, no. 18, pp. 119-122, July 2012.

[2] H. Yao, Y. Sun, and X. Zhang, "Research on no汕neardynamics features of emotional speech," Journal of XidianUniversity(Natural Science), October 2016.

[3] S. Ying, Y. Hui, X. Zhang, and Q. Zhang, "Feature extraction of emotional speech based on chaotic characteristics," Journal of Tianjin University, vol. 48, no. 8, pp. 681-685, August 2015.

[4] Y. E. Jixiang, "Speech emotion recognition based on multifractal," Computer Engineering and Applications, vol. 48, no. 13, pp. 186-189, 2012.

[5] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, "An optimal two stage feature selection for speech emotion recognition using acoustic features," International Journal of Speech Technology, vol. 19, no. 4, pp. 1-11, August 2016.

[6] I. Trabelsi and M. S. Bouhlel, "Feature Selection for GUMI KernelBased SVM in Speech Emotion Recognition;'International Journal of Synthetic Emotions, pp. 57-68, August 2016.

[7] Y. Sun and G. Wen, "Emotion recognition using semi-supervised feature selection with speaker normalization," Springer-Verlag New York, Inc., September 2015.