IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION WITH DEEP LEARNING USING FUSION

S. SREEJA, N. KRISHNA CHETAN, K. SANDHYA, N. MANOHARI
Final Year – Department of Electronics and Communication Engineering.
Jyothishmathi Institute of Technology and Science Karimnagar, Telangana.

ABSTRACT

Since a long time ago, emotion identification from the voice signal has been a research issue in applications involving human-machine interfaces. Many techniques have been developed to extract the emotions from the speech stream. This paper reviews speech emotion recognition basedon earlier technologies that employ several classifiers for emotion recognition. The classifiers are used to distinguish between many emotions, including neutral, happiness, sadness, and angry. Based on retrieved features, classification performance is achieved. In a two-stage process, featureextraction and a classification engine, . First, two sets of features are checked: the first is a 42- dimensional vector of audio features that contains 39 coefficients of the Mel Frequency Cepstral Coefficients (MFCC), the Zero Crossing Rate (ZCR), the Harmonic to Noise Rate (HNR), and the Teager Energy Operator (TEO). And for the second, we suggest using the Auto-Encoder approach to choose relevant parameters from the parameters that had previously been extracted. Second, asa classifier approach, we use Support Vector Machines (SVM) and Random Forest (RF). this paper suggests a speech signal-based emotion recognition system.

1.INTRODUCTION

There are various ways to communicate, but voice is among the quickest and most individual‘s behavioral communication signals. Speech can therefore be a quick and effective way for humans and machines to communicate. Humans naturally possess the capacity to utilize all of their senses in order to fully understand the message that has been received. People are truly able to detect their communication partner's emotional condition through all of their accessible senses. Although emotional recognition comes naturally to humans, it is an extremely challenging assignment for machines.

It is very challenging to detect an emotion from a speaker's words for the following reasons: Which specific speech elements are more helpful in identifying different moods is unclear.

Speech featuresare directly impacted by the variety of phrases, speakers, speaking styles, and speaking rates that have been presented as a result. The same statement might convey a variety of feelings. Each emotion may match

a specific part of the spoken sentence. As a result,it is quite challenging to distinguish between these utterance fragments.

The annotation of speech emotion recognition is challenging, because it is affected by the annotator‘s bias towardlinguistic, cultural, and social constraints. Speech emotion recognition, in particular, yields biased spurious correlations between speaker characteristics and the emotional class of the recording.

Several datasets have been created over the years for the training and evaluation of machine learning methods for speech emotion recognition . Most of them are relatively small. The difficulty in learning from biased and small datasets are the main challenges facing the deep learning research community.The classic self-supervised learning process relies on are presentation trained on a large unlabeled dataset, and a downstream task trained on a relatively small labeled dataset.

Generally, our method enhances a downstream task performance by using a third dataset with labels different from the downstream task labels. For example, in this work, for speaker emotion recognition,our method normalizes undesired characteristics from the self-supervised representation to improve performance on the speech emotion recognition task. We carry this out by learning a feature representation that excels at speech emotion recognition while being robust enough for speaker characteristics.

Another issue is that how emotions are expressed depends on the speaker, as well as his or her surroundings, culture, and upbringing. The speaking style also changes with the culture and surroundings, which presents additional barrier for the speech emotion. In this paper, we firstlysuggested combining other characteristics—the coefficients MFCC, ZCR, and TEO—with a new characteristic, the Harmonic to Noise Rate HNR, in our emotion recognition system.

We combined all the techniques into a single input vector in order to increase the identification rate. Because these techniques are more frequently employed in speech recognition and have high recognition rates, we decided to use the coefficients MFCC, ZCR, and TEO in our work. And secondly, we suggested using an auto-encoder to reduce the input vector dimensions in order to optimize our system. Support vector machines (SVM and random forest (RF) were used.

2. LITERATURE SURVEY

1.**A Review on Emotion Recognition Algorithms Using Speech Analysis- Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Malik Arman Morshidi, Mira Kartiwi.**

The model they used is Convolutional Neural Networks. The architecture contains following layers Convolutional Neural Network, Batch Normalization, ReLU Activation function, Max-Pooling, andsome Dense layers from keras library. To the data available, initially they added noise, stretch time, shift pitch as a way to reduce overfitting is to increase the size of the training data and extract features using any of the following such as Mel Spectrogram, Chroma STFT, MFCC.

Adding noise means that the network is less capable to memorize training data samples because they keep changing all the time, resulting network weights will be smaller and a more robust network is formed because of lower loss. Later the data samples are fed into the model.

2.Ruhul amin khalil1 , edward jones, the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah.

The author described the process in In two stage process,the required features are extracted from the preprocessed speech signal and the selection is made from the extracted features. Such feature extraction and selection is usually based on the analysis of speech signals in the time and frequency domains. During the classification stage, various classifiers such as GMM and HMM, etc. are utilized for classification of these features. Lastly, based on feature classification different emotionsare recognized.

3.Anjali et al. (2020) review speech emotion recognition approaches.

This review cover 2009 to 2018 and several features applied in SER. Despite its limitations, it can still be considered as a starting point. Also, Paruchuri (2015) documented a review on the significance of speech emotion features such as noise reduction and dataset. The importance of diverse classification methods including support vector machine and hidden Markov model. Identification of various features associated with SER was considered the strength of the studywhile leaks of modern approaches were identified as a limitation. Also, they suggested recurrentand convolutional neural networks as a deep learning technique.

4.M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, „„Recognizing emotions induced by affective sounds through heart rate variability,"" IEEE Trans. Affect. Comput., vol. 6, no. 4, pp. 385–394,

Firstly, they utilized the static of log-Mel spectrum, delta and deltas-deltas that are combined to make up the feature vector together from raw speech signal as proposed model‘s input. Then they introduced e dilated convolution and residual unit which is the skip connection trick attached with the tradition convolution networks, and then take advantage of the feature from DRN (dilated CNN with residual block) to be fed into the BiLSTM layer to extract further features, and make them passthrough the attention mechanism at last. In additional, the author optimized the loss function to use the center loss which helps our model distinguish the features more easily. According to the experiment they conducted ,the author got the better results compared with ACRNN model and other previous works in the area of SER.

3. PROPOSED SYSTEM

The proposed system consists of classifiers that are used to distinguish between many emotions, including neutral, happiness, sadness, and angry. Based on retrieved features, classification performance is achieved. In a two-stage process, feature extraction and a classification engine, .

It is believed that a proper selection of features can have a significant impact on the classification performance. Many diverse audio features are assessed in the literature to boost up recognition ratein SER. However, necessarily not all of them have a positive impact on emotion recognition. In fact, having too many features can reduce the performance and/or increase the computing time. As a result, only a set of most significant features is considered in this study. Pitch, Energy and Intensity are traditional but important prosodic features of speech which provide valuable

information to differentiate emotional states.

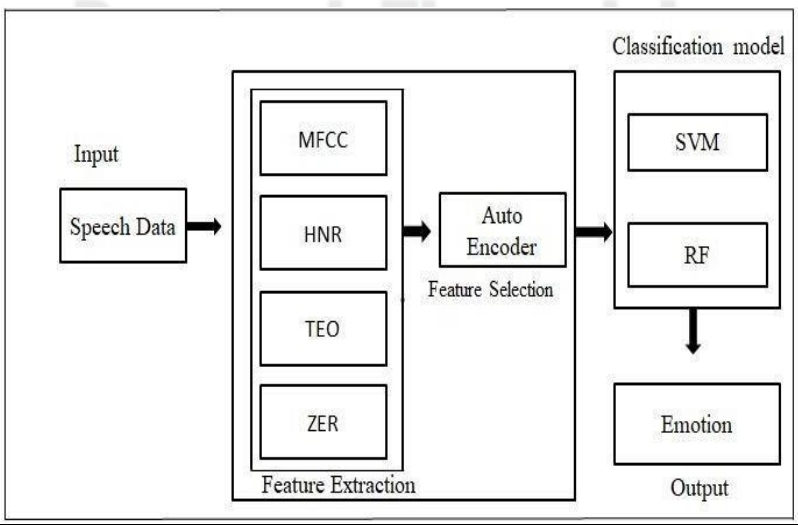
The resonant frequencies, on the other hand, are produced in the vocal tract referred to as formants in several forms, each at a different frequency, occurring at roughly 1000Hz intervals. The first three formants convey valuable information and are used in our experiments. Mel-frequency cepstral coefficients (MFCCs) are other common features that are used in fields like speech and gender recognition, music information retrieval and recently used extensively in SER. The 12 MFCC coefficients are used in our experiments. Forpitch contour configuration, the standard range 75 to 500 hertz is considered, which means that the pitch analysis method will only find values between 75 and 500 Hz. Using Praat software [29], 68 sound or speech features are extracted from utterances representing information such as Duration, Pitch, Intensity, the first three Formants, Amplitude, Harmonicity or Harmonics-to-Noise Ratio (HNR), Jitter, Shimmer, Energy, Energy-Air, Power, Zero-cross-rate (ZCR) and the first 12 MFCCs. Table 3 indicates the set of 68 acoustic features which form our baseline

feature set. As it is shown in table 3, we adopt several statistical parameters such as Minimum, Maximum, Median, Root-Mean-Square and standard deviation of explained features. Next section explains how applying various feature selection algorithms could impact the accuracy of the system.

First, two sets of features are checked: the first is a 42-dimensional vector of audio features that contains 39 coefficients of the Mel Frequency Cepstral Coefficients (MFCC), the Zero Crossing Rate (ZCR), the Harmonic to Noise Rate (HNR), and the Teager Energy Operator (TEO). Because thesetechniques are more frequently employed in speech recognition and have high recognition rates,we decided to use the coefficients MFCC, ZCR, HNR , and TEO in our work. And for the second, we suggest using the Auto-Encoder approach to choose relevant parameters from the parameters that had previously been extracted. We propose the use of the auto-encoder to reduce the numberof features and to compare the results ofclassification with the systems use 42 features.

The main reason behind using feature selection is to eliminate features which are irreverent and redundant. Advantages of using feature selection algorithms are considered twofold: training time could significantly be reduced and it also helps minimize the problem of overfitting. In this study, in order to form our ―golden set‖ of sound features, two powerful feature selection methods, SVMattribute evaluation and Correlation-based Feature Subset Selection (CFS) are applied to theextracted feature sets and results are compared to each other.

We have started to modify the parameters of the basic AE in order to give better identification rate by the



use of the RBF kernel of SVM classifier.Second, as a classifier approach, we use Support Vector Machines (SVM) and Random Forest (RF).The proposed architecture of our work is shownin below figure.

4. IMPLEMENTATION

4.1 Input and Output

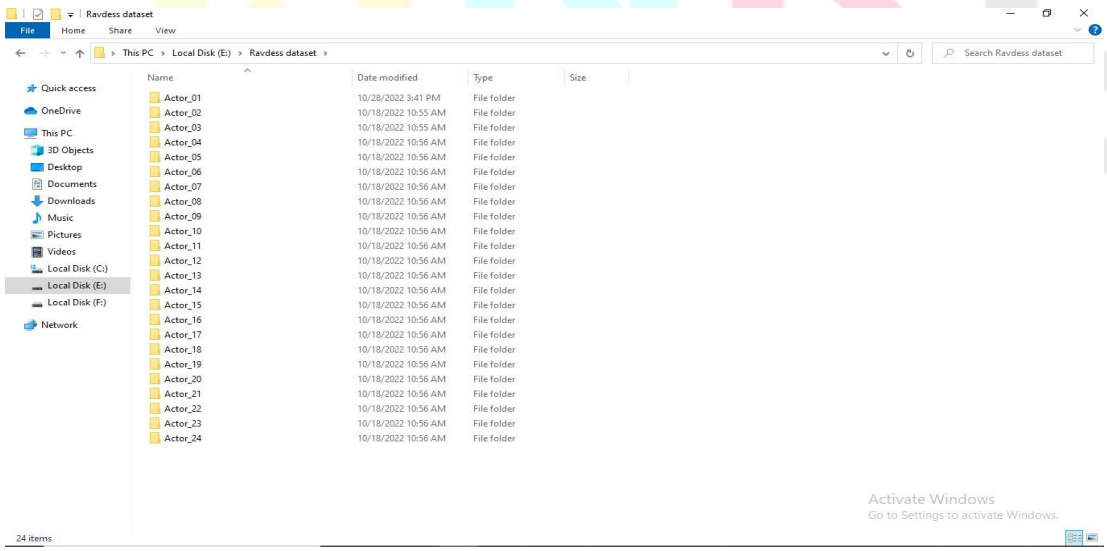
Steps for implementing project

- Step1: The feature extraction procedure is essential for speech emotion recognition. The accuracy of the classification results is directly influenced by the features' quality.The feature extraction process takes 4 parameters that are MFCC,HNR,TEO,ZER that are combined in a single vector.
- Step2: Ravdess Dataset supports the following emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised. The emotions that we are going to observe are neutral, sad, happy and angry.
- Step3: The desirable features are selected by using auto encoder(features that are extracted fromthe vector). This paper selects auto-encoder (AE) in feature selection.
- Step4: An auto-encoder has a number of parameters, including the number of hidden layers, the unit in each layer, weight regularisation parameter, and the number of iterations. The new reconstructed data (output of AE) has been reclassified as training data in order to train the combination of SVM and RF model to predict test samples.
- Step5:The fusion of Support vector machine and Random Forest is used to classify the emotion and predicting the emotion in the speech.
- Step6: In Ravdess Dataset, for each path, get the basename of the file, the emotion by splitting thename around _-‘ and extracting the third value in the path file. Using our emotions dictionary, this number is turned into an emotion, and our function checks whether this emotion is in our list of observed_emotions; if not, it continues to the next file.

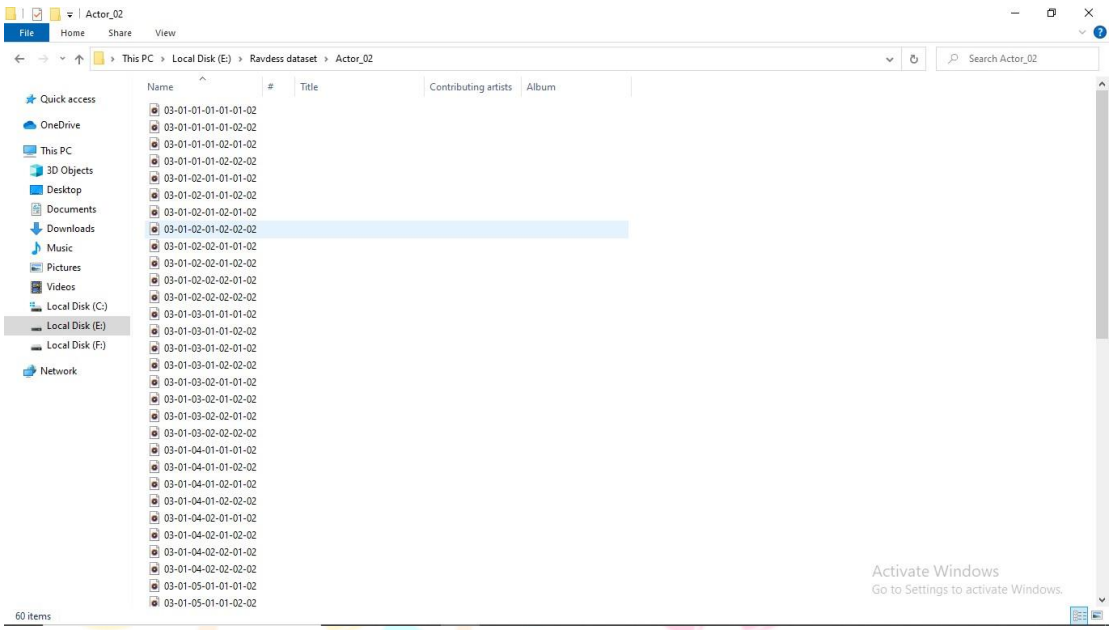
Then our model will be giving output as a text which represents the emotion in the speech.

Input Design

The input for this implementation of the project consists of audio files.



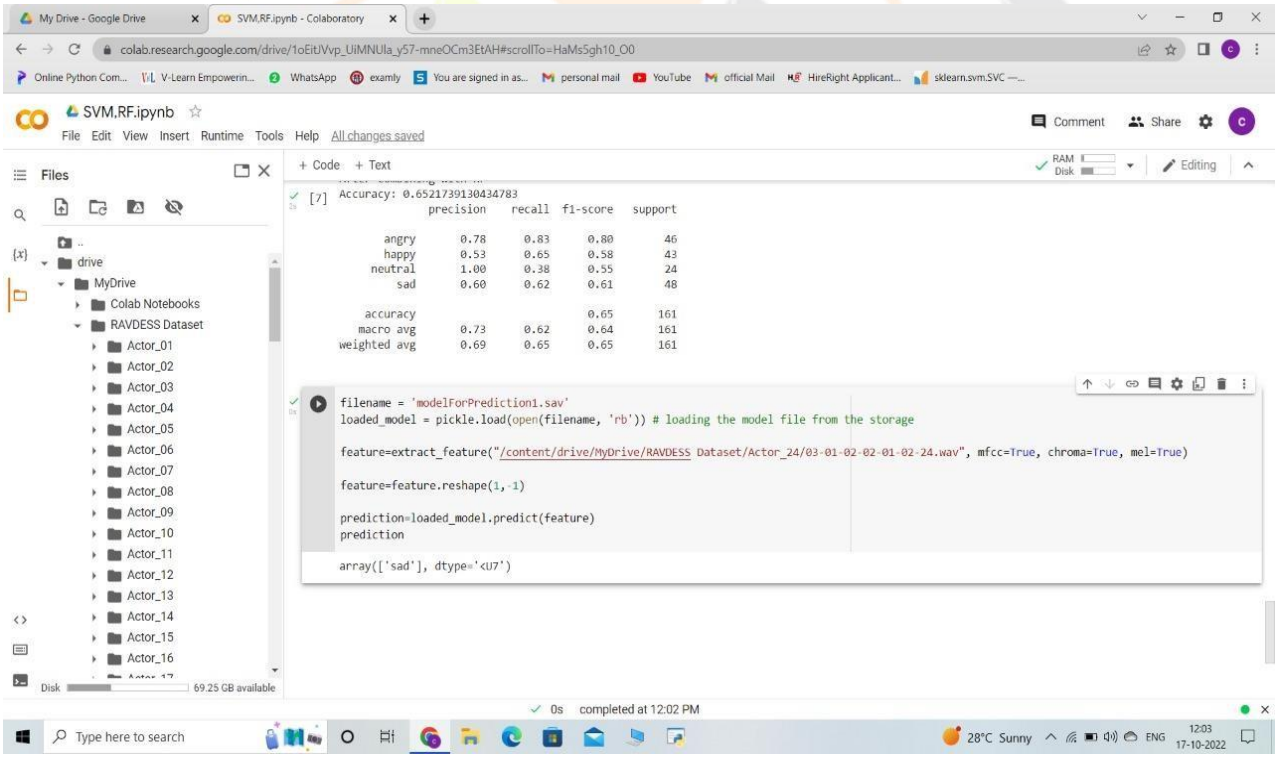
Screenshot of folders in Dataset



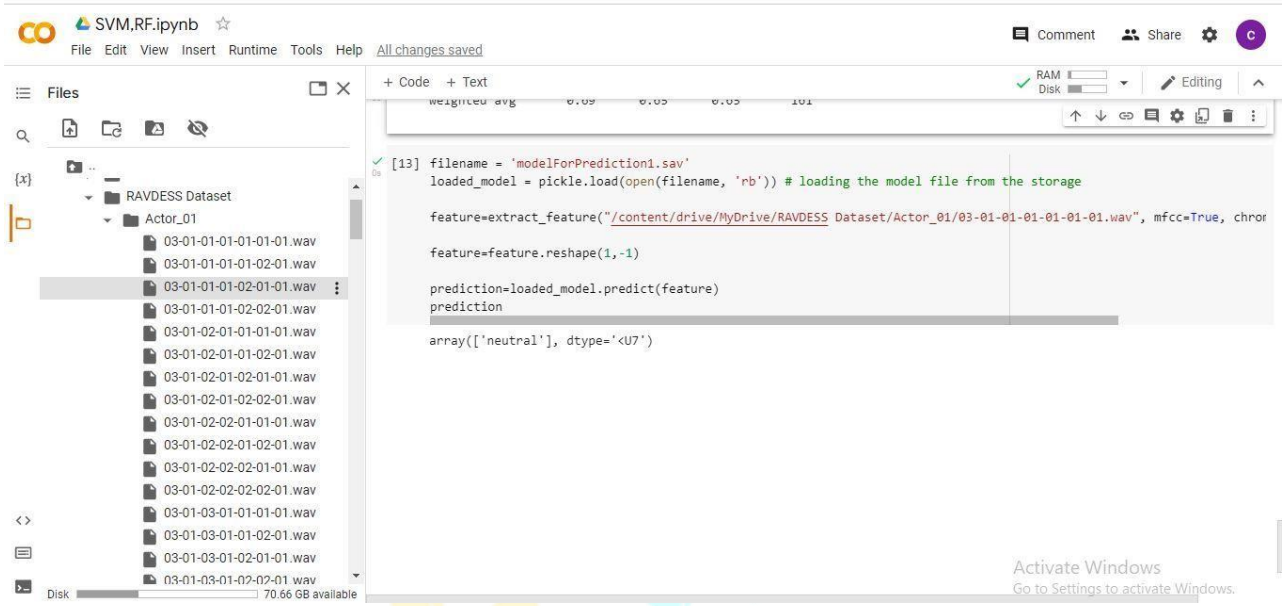
Screenshot of Audio files in Dataset

Output Design

The output design of the implementation consists of the Emotion‘s name, the location where he is identified.



Output Screen shot(Sad Emotion)



Output Screen shot(Neutral Emotion) The accuracy & classification of emotions depicted

```
After combining with RF
Accuracy: 0.6521739130434783
              precision    recall  f1-score   support

       angry       0.78      0.83      0.80        46
       happy       0.53      0.65      0.58        43
     neutral       1.00      0.38      0.55        24
         sad       0.60      0.62      0.61        48

    accuracy                           0.65       161
   macro avg       0.73      0.62      0.64       161
weighted avg       0.69      0.65      0.65       161
```

Accuracy

WORKING: The working principle of this system can be broken down into several steps

Data Collection: The first step involves collecting a dataset of speech samples that are labeled with corresponding emotions. This dataset is crucial for training the deep learning model.

Feature Extraction: Once the dataset is available, the system performs feature extraction on the speech samples. This involves converting the raw audio signals into a representation that can be used by the deep learning model. Popular feature extraction techniques include Mel Frequency Cepstral Coefficients (MFCCs), which capture the spectral characteristics of the speech.

Deep Learning Models: The system utilizes deep learning models to learn the underlying patterns and relationships between the extracted features and the emotional content of the speech. These models are typically neural networks with multiple hidden layers that can automatically learn hierarchical representations of the input data.

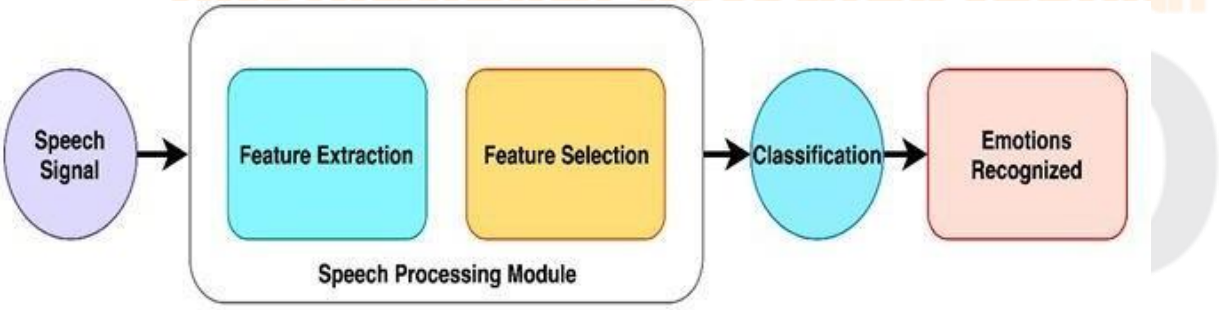
Fusion Techniques: Fusion refers to combining information from multiple sources to make a more informed decision. In the context of speech emotion recognition, fusion can be performed at different levels, such as feature-level fusion, decision-level fusion, or model-level fusion. This step involves combining the outputs of multiple deep learning models or different feature sets to enhance the overall recognition performance.

Training: The deep learning models are trained using the labeled dataset. The training process involves feeding the extracted features into the models and adjusting the model's parameters to minimize the difference between the predicted emotions and the ground truth labels. This is typically done using optimization algorithms like stochastic gradient descent.

Evaluation and Testing: After the models are trained, they are evaluated using a separate validation set to measure their performance. Various metrics such as accuracy, precision, recall, and F1 score can be used to assess the effectiveness of the models. Once the models are deemed satisfactory, they can be deployed for real-time emotion recognition.

Real-Time Emotion Recognition: In the real-time scenario, the system processes incoming speech signals and extracts the relevant features using the trained models. The fusion techniques are applied to combine the outputs of multiple models or feature sets, leading to a more reliable and accurate emotion recognition result.

By combining deep learning techniques and fusion methods, the Speech Emotion Recognition system can effectively identify and classify emotions expressed in human speech. It has the potential for various applications, including human-computer interaction, call center analytics, and emotion-aware systems.



RESULT

Efficiency

In our work, Four basic human emotions are expressed: Angry, Fear, Happy, Sad, and Neutral. We use only the language English the number of audio become 241 audiovisual samples (75% dedicated for learning and 25% for testing). The best results for different emotions using SVM and RF are summarized in the following Table

Emotions	39 MFCC, ZCR, TEO, HNR without feature selection
Angry	75.00
Sad	71.42
Happy	72.72
Neutral	80.32

The recognition rates on the test corpus obtained

We propose the use of the auto-encoder to reduce the number of features and to compare the results ofclassification with the systems use 42 features.

The parameter number of unit in hidden layer is fixed at 35 units, we vary the parameter number of iteration to have a maximum identification rate (we found a rate equal to 71.60 % when number of iteration = 10000). After fixing the two previously parameters we have varied the parameter the weight regularization parameter and which gives a better identification rate inthe order of 72.83% when weight regularization parameter= 0.00001.

The table below summarizes the best recognition rates found for the four emotions using basic AE as feature selection with SVM and RF.

Emotions	Basic AE with 35 units in hidden layer
Angry	83.33
Sad	78.57
Happy	81.81
Neutral	85.21

The best results obtained for the six emotions using the basic AE

CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

In order to first identify emotion with SVM and RF, we showed the performance of our suggested systems in this research. These systems combine the HNR feature with the three extensively used features in emotion recognition (MFCC, ZCR, and TEO). Second, we present our proposedmethod of implementing our system,

which involves using the auto- encoder dimension to condense the features that had previously been recovered using combined vector.

When compared to previous study emotion recognition systems, the outcomes of these systems demonstrate their usefulness in producing positive results. We demonstrate how the use of auto- encoder dimension reduction increases the rate of identification.

FUTURE ENHANCEMENT

In future, we can consider employing different feature types, applying our system to larger data bases, and using different feature dimension reduction techniques. Last but not least, we can also think about recognizing emotions using an audiovisual foundation, in which case we can take advantage of descriptors from speech and others from image. This enables us to raise the rate at which each emotion is recognized.

REFERENCES

[1].A.Mehriban –Communication without words‖, Psychology Today, cilt 2, no. 4, pp. 53-56.,2018

[2].B. Fasel ve J. Luettin, –Automatic Facial Expression Analysis: A Survey‖, Pattern Recognition, cilt 36, pp. 259-275, 2003.

[3] N.Umapathi., N.Ramaraj, (2014) Swarm Intelligence based dynamic source routing for improved quality of service, Journal of Theoretical and Applied Information Technology, Pp 604-608.

[4] N.Umapathi., elt .,(2017), "Optimizing link quality and bandwidth estimation for dynamic source routing". **IEEE** international conference on algorithms, methods, model and application in emerging technology, at Bharath institute of higher education and research, held on 16th to 18th feb 2017. 10.1109/ICAMMAET.2017.8186636

[5].Rao K. S., Kumar T. P., Anusha K., Leela B., Bhavana I. And Gowtham S.V.S.K., –Emotion Recognition from Speech‖ (IJCSIT)International Journal of Computer Science and Information Technologies, Vol. 3 (2), pages: 3603-3607,2012

[6].Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A.Empathyand emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1), 98-105,2013

[7]. Polat, G., & Altun, H. Ses Öznitelik Vektörlerinin Duygusal Durum Sınıflandırılmasında Kullanımı.Eleco.,2016

[8]. P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, "Speech emotion recognition using kernel sparse representation based classifier," in 2016 24th European Signal Processing Conference (EUSIPCO), pp. 374-377, 2016.

[9]. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. Audio-visualemotion recognition in video clips. IEEE Transactionson Affective Computing,2017

[10]. S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusionof teager energy operator and MFCC," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.

[11].S.B. Reddy, T. Kishore Kumar , "Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features," in IEEE 18th International Conference on Advanced Learning Technologies, 2018.

[12]. Zamil, Adib Ashfaq A., et al. "Emotion Detection from Speech Signals using VotingMechanism on Classified Frames." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.

[13].L.X.Hùng : Détection des émotions dans des énoncés audio multilingues. Institut polytechnique de Grenoble, 2019.

[14]. Ferrand, C: Speech science: An integrated approach to theory and clinical practice. Boston, MA: Pearson, 2017.