ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Novel approach of Multiclass Speech emotion classification by density clustering with convolution neural network Features with SVM

[1]Krishnaiah Nallam, [2]Dr. P. Santosh Kumar Patra, [3]Regonda Nagaraju
[1]Professor, Dept of IT, St.Martin's Engineering College, Secunderabad, Telangana,India.
nkrishna520@gmail.com Orcid: 0000-0001-6486-7170
[2]Principal & Professor, Dept of CSE, St.Martin's Engineering College,
Secunderabad, Telangana, India.
drpskpatra@gmail.com
[3]HoD & Professor, Dept of IT, St.Martin's Engineering College, Secunderabad,
Telangana India.
nagcse01@gmail.com

Abstract : In recent studies conclude that speech contains rich information about emotion.so using this information to understand the human interactions. Classification of emotions by speech information is challenging because analysis the pattern of features will change human to human and speaker to speaker. The proposed approach use Deep Learning specific architecture CNN (Convolution Neural Network) and clustering by Dense stream Clustering. The proposed approach use transfer features approach which uses CNN layers for extracting different features from different layers because different layers generate dynamic features pattern due to its activation function. We Use these features in SVM (Support Vector Machine) for learning because it demonstrate the difference between Na Yang et al. and our approach.

Keywords: SVM , Emotion, CNN, Dense clustering

1 Introduction

Our communication way is developing with the advancement in technology. As we are going to study about recognition of speech emotion by several techniques, first of all we discuss about the speech. Speech is vocalizes form of the communication used by human beings. There are various consonant and vowels used in the vocabulary set for communication purpose. In speech production our thoughts are translated to words by choosing different vowels and consonants. Speech is nothing but a form of sound used in vocal language which comes from lungs to the vocal cord which floats in the air. Human express its emotion by using a set of vocabulary. With the passage of time and development in technology human created various ways to express their emotion and also for communication. As we can easily understand the emotion of an individual with the help of speech many researchers are used to classify the emotion on the basis of speech. Ekman gives a classification on emotions whichincludes

Analysis of emotion expressed in voice can be done in three different levels:

1) **Physiological level:** It studies body response to any behaviors or activity in an organism. In this level main focus on the brain cell, structure, component and chemical interaction involved to produce action. Their attentions are generally on topics like sleep, emotion, ingestion, sense etc. It describes the nerve impulse pattern of majorstructure.
2) **Phonatory-articulation level:** Position and movement of structure are studied
3) **Acoustic level:** Characteristics of speech are described in this level. It gives attention on the waves coming frommouth.

It is very important to recognize the speech and emotion of speech. While using any machine or robot it becomes necessary to recognize the emotions of speech. Now a day in security system we feed our voice in that case too we

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

need an accurate recognition system. There are various methodologies to assess the emotion or speech. Trained Observational Coders are used to classify the emotions. As it is a time consuming and costlier method therefore we needed improved methods for the classification of emotions with high accuracy rate.

Speech emotion Classification is a process of analyzing the vocal behavior of the persons. It works on the identification of an individual person with help of the voice. Human speech contains a lot of information for conveying the message related to their needs, desire, and thoughts. The richness of human speech for understanding the emotions motivated the researcher to enhance the field of emotion classification. In speech emotion classification, a large number of low level and high level features has been devised to isolate information of emotions in speech messages its include pitch of voice message, energy, timing and quality of voice. There are various fields where the speech emotion classification technique is used for the identificationpurpose.

- Identification insmartphones.
- Identification in BankTransaction.
- For automatic control invehicles
- For high performance in fighter aircrafts.

Advantages of Speech Classification System

- We can detect the human emotions and feelings even if we cannot able to understand the language of that person. It helps in the communicationprocess.
- Voice signals are easy to record even if the weather is in extreme condition like high humidity, high light and hightemperature.
- No special training is required to use thistechnology.
- It is very useful for the peoples with disabilities, they can use their voice for command thesystem.

2 Related Work

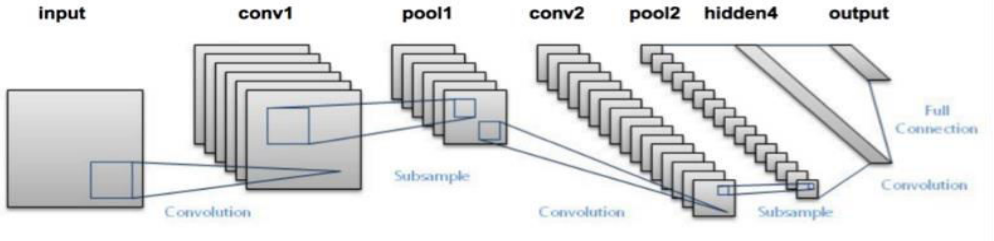
Thresholding Fusion mechanism is used to combine the output of various SVM that increase the accuracy rate of the system. System performance is evaluated on the noisy speech signal, non - professional recordings and speaker tests [1].Speech recognition model using Hidden Markov process is used to recognize emotions. Two methods were compared in this paper. The first method is based on Gaussian mixture model which identify the raw pitch and speech signal. Second method based on temporal complexity using continuous Markov model. This method works with the German and English language speakers. This method gives better recognition rate [2].A speech emotion recognition system that is totally Acoustic Prosodic (AP) and Semantic Labels (SLs) based on multiple classifying systems. The basic classification in this system is three models, GMMs, SVMs and MLPs. For the recognition of emotions, the maximum entropy model (MaxEnt) is used to identify the relationship between emotional and EAR states. For combining the AP and SL recognition with the weighted product fusion method.[3].The multimodal automatic emotion recognition method is used for speech emotion recognition. A database was used in which speech sample of different people is stored. In this paper unimodal data, bimodal and multimodal data is used and based on the Bayesian classification. After the automatic classification of each modality, all were combined using multimodal modality. Multimodality giver better result when compared with other modality [5].Sparse auto-encoder based feature learning method is used to find common feature in small target data and apply this to the reconstruct source data to complete the transfer of knowledge data from source to target. Linear kernel with SVM classifier is used for the classification process [6].The difference between the first order and second order harmony features plays a vital role in the speech recognition. Fourier parameter method is used for identifying the

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

quality of content and harmony features for speech emotion recognition. Fourier parameter (FP) and Mel frequency cepstral coefficient (MFCC) features gives better results[7].A deep learning approach for EEG - based emotional recognition to incorporate spatial electrode information and cross-electrode relationship within CNNs. We choose to use CNNs because they have the capacity to consider the two - dimensional spatial information[15] .

CONVOLUTION

Convolutional Neural Networks (CNN), were first introduced by Yann LeCun's in 1998 for Optical Character Recognition (OCR), where they have shown impressive performance on character recognition. CNN is not just used for image related tasks, they are also commonly used for signals and language recognition, audio spectrograms, video, and volumetric images.



CNN uses multiple layers in its architecture. Following are the layers used to build convolutional neural network architectures.

Convolutional Layer
Activation Layer
Pooling Layer
Fully-Connected Layer or Densely Connected Layer
Output Layer or Softmax Layer for classification

CNN architecture is explained in detail in section 3.

Convolutional Neural Networks (CNN), were first introduced by Yann LeCun's in 1998 for Optical Character Recognition (OCR), where they have shown impressive performance on character recognition. CNN is not just used for image related tasks, they are also commonly used for signals and language recognition, audio spectrograms, video, and volumetric images.

Convolutional Layer

Convolution Layer provides a convolution operation, in which a 2-D or 3-D filter of appropriate size sweeps over an image and apply the filters to each depth of an image. The convolutional layers are restricted version of the Multi-Layer Perceptron (MLP) adapted to take a 2D / 3D inputs instead of 1D. The idea behind convolutional layers is to detect elementary features such as edges, corners, and endpoints, and combine them using multiple layers to get high-level features that might describe an object completely.

Moreover, this architecture is designed for high-level features extraction from an image at any given layer to describe an object like face, chair, or a car. In addition to this, convolution also provides an important and valuable feature attribute called shift invariance. That is, if the input to the first layer is shifted, then the output of the first layer is also shifted by the same amount. Convolution has 2 main parameters which can change the behavior of convolution, like stride and padding.

Output of the convolution layer is calculated as per the following formula.

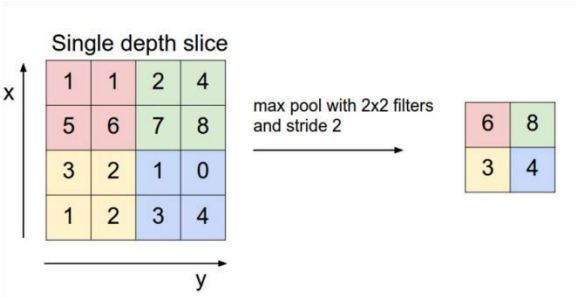
$$W_{new} = \frac{(W_{old} - F_{width} + 2P)}{S} + 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

F_{width}: Filter or Kernel size as in width and height parameter while using respective formula.
P: Padding
S: Stride window size for convolution
W_{new}: New width of the output image
W_{old}: Old width of the input image

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Pooling Layer

Pooling is a method of reducing the feature size in width and height of an input. The pooling operation sweeps a rectangular window over the input feature and computes a size reduction operation for each window (average, max, or max with arg max). Each pooling operation uses rectangular windows of size k, separated by offset strides. For example, if strides are all ones every window is used, if strides are all twos every alternative window is used in each dimension. a simple way of reducing the precision for the position from where distinctive features are located in the feature map. Since the exact position of the feature is irrelevant, only its position in relation to the other features is of importance, especially for classification tasks.



Pooling layer output is calculated as per the following formula.

$$W_{new} = \frac{(W_{old}-F)}{S} + 1 \text{.....(2)}$$

W_{new}: New Width for the output image
W_{old}: Input image width
F: Filter Width size
S: Stride size

This formula is used for the output image width calculation, and same can be used to calculate the resulting height of an output image from the pooling layer by changing width parameter with the height parameter.

Fully Connected Layer

Fully connected layer is the called as dense layer, where each neuron in one layer is connected to each and every neuron in the following layer. This principle is same as the traditional multi-layer perceptron neural network model and how it works. Fully-connected layers refer to be the final layers in the full CNN model. Fully-connected layers operate as a Multi-Layer Perceptron (MLP) with normally either two or three hidden layers and one classification layer. The properties of the MLP make it a superb function approximation, with only two hidden layers it can approximate any function assuming it has enough hidden neurons. Normally, the number of neurons in the hidden layers is constant, with 4096 being a common number for deep networks with large input images.

The fully connected layer can be flattened and connect to the output layer and so on it get reduced in size for the classification of the images. At the fully connected layer, if the input is coming from convolution layer or max pool layer of size XxYxZ size, we can choose how many nodes do we need in the fully connected flattened layer. It could be the XxYxZ number of nodes or (X*Y*Z)/2 number of nodes and then the output will be reduced feed to the output classification layer.

Activation Function

The activation function is really important to the deep neural network, which is complicated and complex. They bring non-linearity property to neural networks. The main property of an activation functions is to convert an input signal to output signal. This is used in every node of the deep neural network for abstraction representation of action potential firing the node.If we don't use the activation function, the output mapping function will be, by default a linear function, which linearity is less effective towards learning of complex function boundaries of the input data. Following are some of the activation functions explained in detail.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Strides

Stride is a concept, which controls the movement the kernel over an image in convolution and max pool operation. By Default, the kernel moves over the image by shifting one position at a time in horizontally or vertically. Starts at (0,0) position of the image and if the stride is 1x1 then it will move 1 in both the direction, horizontally or vertically. Stride is defined as [1,2,2,1], that means, each element in the array is defined as [batch, shift in vertical direction, shift in horizontal direction, channels] respectively.

For an example

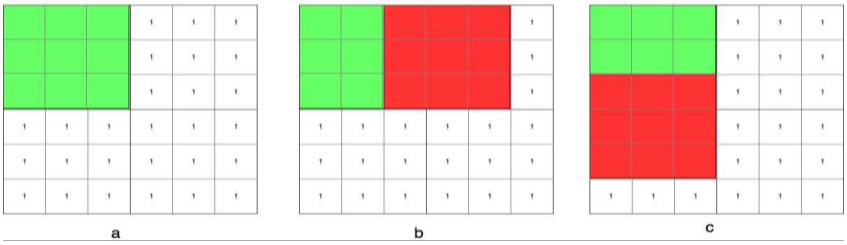


Figure 2: Stride of 2x2 is used for kernel in the convolution process on 6x6 image size

In above convolution, the stride is used as 2x2, in the figure 13b, the kernel moved by 2 positions as shown in red color in horizontal direction and in figure 13c, the kernel moved by 2 positions in vertical direction.

Padding

Padding the input image is 3D or 2D with zeros, such that the convolution layer does not alter the spatial dimensions of the input image. With the zero padding while convolution controls the spatial size of the output image from convolution layer.

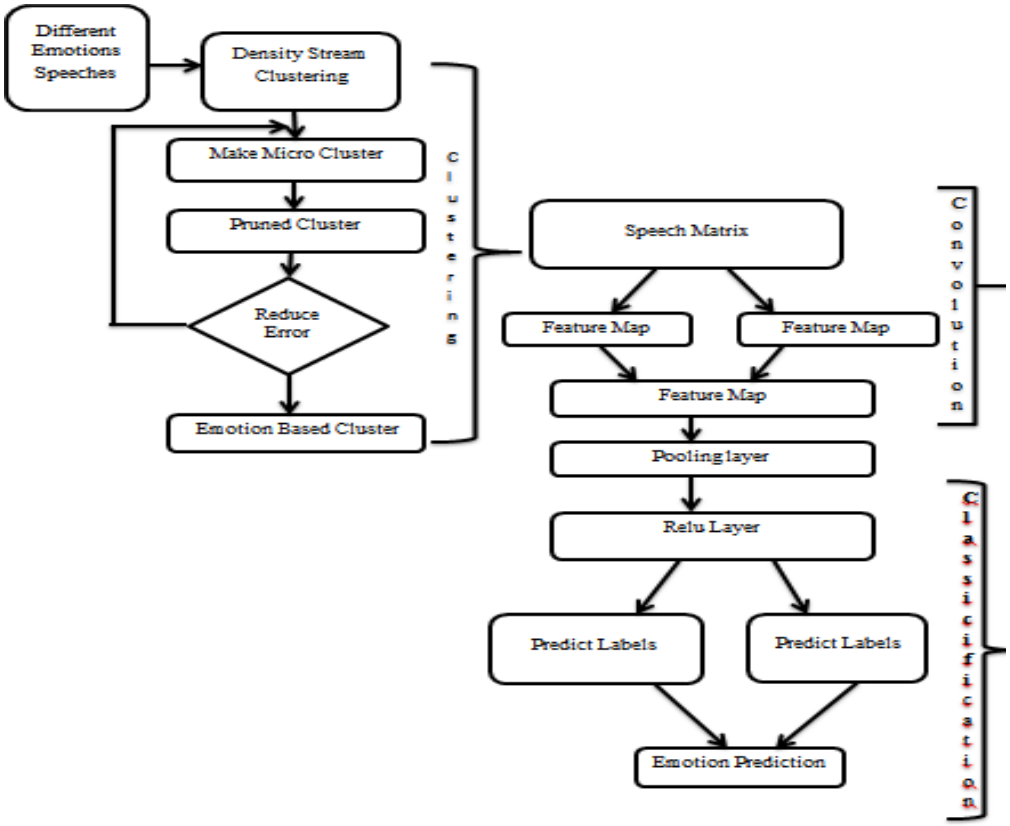
Padding can be calculated as the P = (F-1)/2.

F: Filter size

For an Example convolution without padding or padding = 0

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Convolution Clustering Base Support Vector



3.1 figure 3: Flow Chart of Proposed Methodology

CONVOLUTION

Different CNN structures are used in CNN (Convolution Neural Network). It generated basic structure like Convolution, Pooling and ReLu Layer.

The Convolution layer by

$$E = \sum_{i=1}^{K} \text{II} x - x_i \text{II} \ldots\ldots (3)$$

X is the input, W shared parameter with signal size I*J. ReLu layer Maximize non-linear activation.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

$$(x) = \sum_{c_j \in c_i} w_i \exp\left(-\frac{\text{II}x - c_i\text{II}^2}{2\sigma_i^2}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

X

Third is important layer which learn the features which come from equation(1)

$g_i(x)$

* ( )+ ( )(

S is the simple index: x, p and q they show the classifier with using stochastic gradient descent learning and change W.

Dense Stream Clustering Algorithm
Construct speech cluster(Speech segment: S), Current Clusters : $X_1, X_2 \ldots\ldots X_k$ (Number of cluster: K) K= ~~N~~umber of Emotion **Begin** For each online speaker label J H_J=(0,0,..... 0); For the next speech point in S do Begin Find Closest micro- cluster ; Find the distribution of ADD Micro-clusters change according to Ω Purning the micro-cluster if Purned cluster = K and find statics for Update the number of clusters in(where i is the Pitch in Signal) End End

3.2Convolution Clustering with support vector Classifiction

Speech processing: To obtain a better classification of speech or audio emotion classification by using EMO-db database with sample rate 48 KHz with five emotions. In proposed two speaker source single and doublespeakers.completion of first phase all the objects are placed in some clusters. In second phase average of the early formed clusters is taken. This iterative procedure persists repeatedly until the decisive function given in equation (3) to become the minimum [5]. Supposing that the target object is x, xi designates the average of cluster Ci , decisive function is defined as follows [13].E is the sum of the squared error of all instances in database.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

$$F(x) = \sum_{i=1}^{K} w_i \exp\left(-\frac{\text{II}x - x_i\text{II}^2}{2\sigma_i^2}\right) \ldots\ldots (6)$$

Feature Transfer and Feature Generation: Different layer of CNN features are transferred C_1-C_5 for classified in three class then C_5 layer features for five classesclassification.

Target Classifier (SVM): we use standard SVM classifier which use Transferlearning. Kernel parameters, proposing to increase the class-based separability of the data in given feature space. Now for or prediction of an instance at x with unknown class value then the likelihood function, Lj (x) must be maximum to represent a specific class as shown in Equation(6). As we have sufficient instances, then by using law of large numbers, probability density function at class j can be given by Equation(5). Here $D(x_i)$ is maximum distance between xi and ki nearest instances of same class. Γ is Gamma function [17].

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

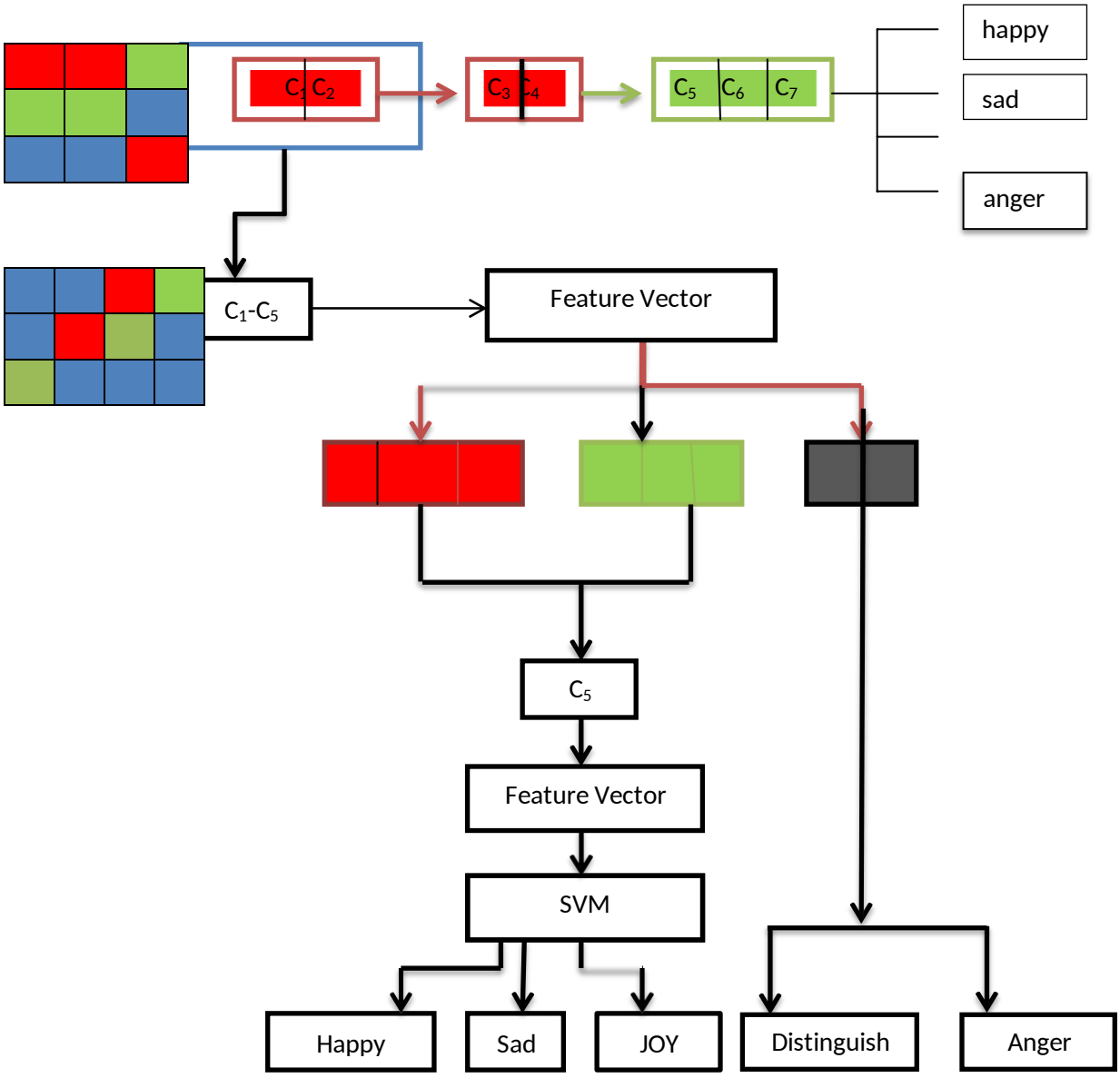


Figure 4: Classification Procedure with features using CNN

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Results

Table 1.1 Single Speaker Results

Single Speaker	Existing (Precision)	Proposed (Precision)	Existing (Recall)	Proposed (Recall)	Existing (Accuracy)	Proposed (Accuracy)	Existing (F-measure)	Proposed (F-measure)
EMO-DB	86.34	88	87.45	89	87.34	90.23	86	88.49
SAVE E	85.34	86.23	85.45	87.23	86.56	88.43	85.88	86.72

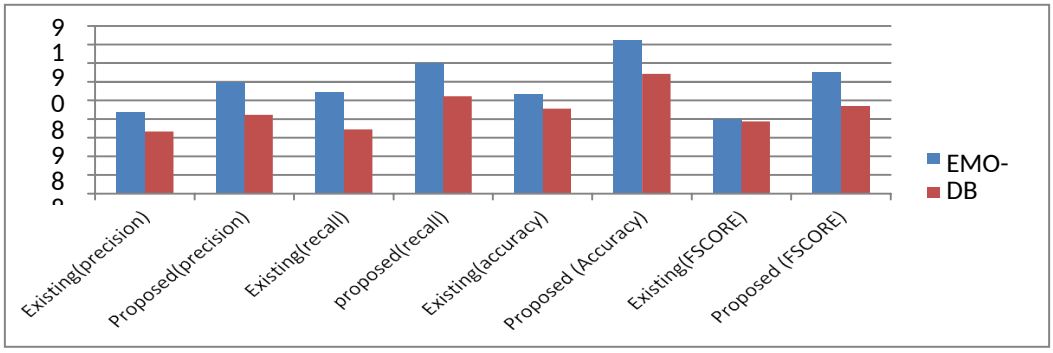


Figure 5: Graph of single speaker results.

Observation 1: Table 1.1 show the first experiment results which is proposed between (CNN and Clustering) Existing (SVM) based approach on single speaker. In comparison, existing approach show the reduction of False Positive Rate because dynamic threshold is given by different layers of CNN. In experiment working on EMO-DB and SAVE E data set, both the data set significantly improved in proposed approach because feature transfer of clustering signal to normal database.

Table 1.2 Dual Speaker Results

Dual Speaker	Existing (Precision)	Proposed (Precision)	Existing (Recall)	Proposed (Recall)	Existing (Accuracy)	Proposed (Accuracy)	Existing (F-measure)	Proposed (F-measure)
EMO-DB	85.34	87	86.34	87.23	88.34	89.23	86.34	87.11
SAVE E	84.34	86	87.12	88.23	84.23	85.23	85.3	87.1

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

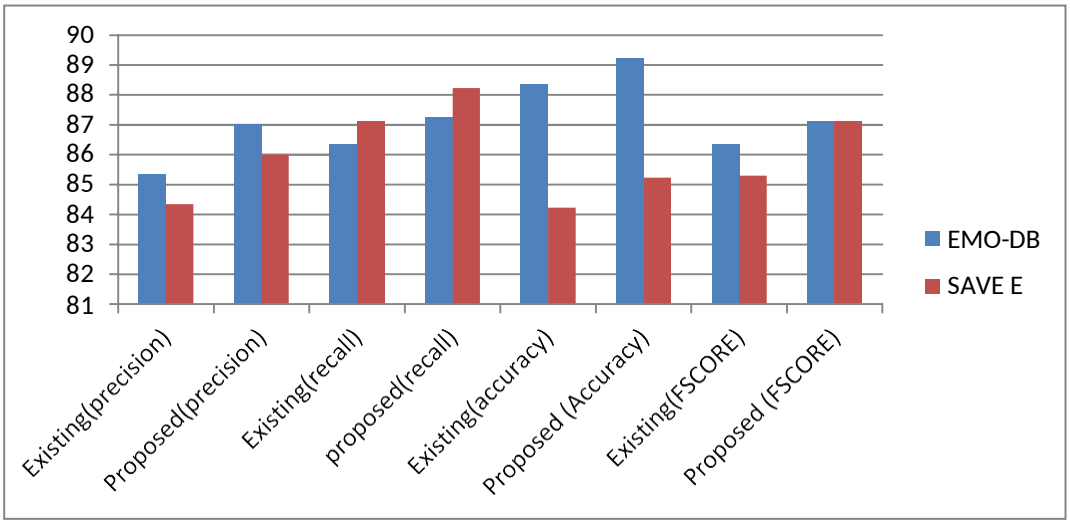


Figure 6: Graph of Dual speaker results.

Observation 2: Table 1.2 depicts the comparison between emotions in speech signals. In this experiment, two speakers of same and different voice which is mixed in cluster and apply the dense stream clustering and after this apply CNN layers for features and learning by SVM (Support Vector Machine)

This experiment work on SAVE E and EMO DB dataset, in data set accuracy, precision and recall is improved to existing approach.

Table 1.3 Comparison results of EMO-DB and SAVE E

Emotions	Database	Precision	Recall	Accuracy	F-measure
Happy	EMO-DB	70%	75%	72%	72%
	SAVE E	82%	87%	88%	84%
Sad	EMO-DB	82%	85%	90%	83%
	SAVE E	87%	88%	87%	87%
Joy	EMO-DB	78%	87%	95%	75%
	SAVE E	86%	88%	90%	86%
Angry	EMO-DB	88%	89%	87%	88%
	SAVE E	87%	88%	86%	88%
Distinguish	EMO-DB	95%	92%	83%	88%
	SAVE E	89%	89%	89%	89%

Observation 3: Table 1.3 shows the analysis of individual emotion classification pattern and take average of both experiment 1 and 2 in proposed approach. Above discussion prove that the proposed approach does well when it compared with the existing approach, so we analyzed the Individual signal of emotion. The analysis shows that distinguish emotion highly identified as compare to other emotions. So we can say that we are easily classified the emotions from the speech.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

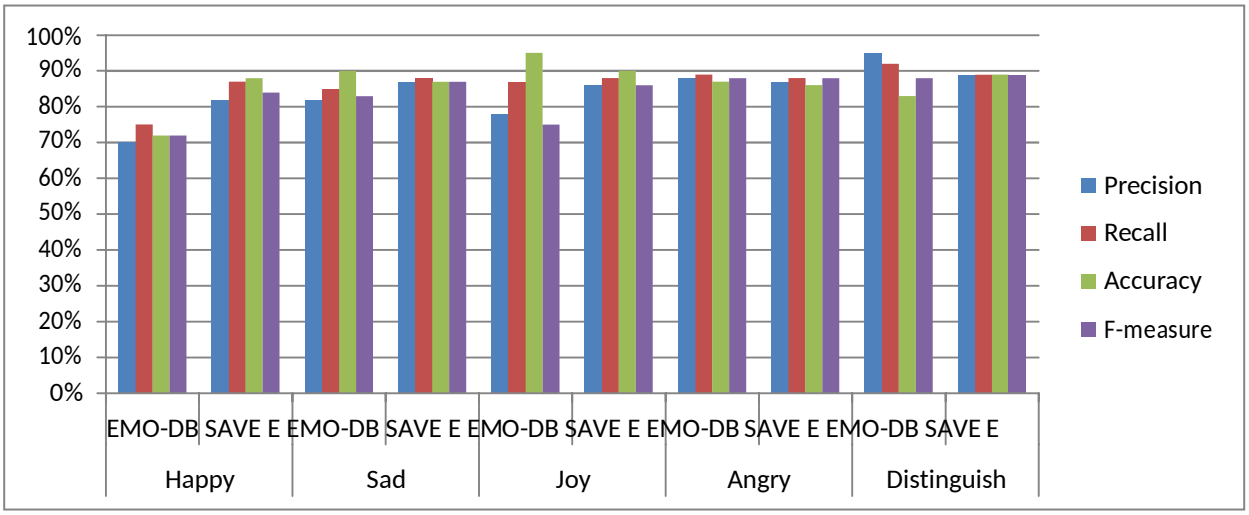


Figure 7: Comparison Graph of emotions speaker results.

Conclusion

This paper presents the multi-class classification by transfer the features from different layers of CNN and then learns by using multiclass SVM. Core of this research is that we have not changed and use any static threshold of parameters. We also cluster the voice or speech of different speaker then cluster by Density-based clustering which makes dynamic clusters, but constraint put on cluster equal to the number of emotions in speech. Our results significantly improve to SVM with the static threshold to slack parameters.

References

[1] Yang, Na, et al. "Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification." International Journal of Speech Technology 20.1 (2017): 27-41.
[2] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. Vol. 1. IEEE, 2003.
[3] Wu, Chung-Hsien, and Wei-Bin Liang. "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels." IEEE Transactions on Affective Computing 2.1 (2011): 10-21.
[4] Nwe, Tin Lay, Foo Say Wei, and Liyanage C. De Silva. "Speech based emotion classification." TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. Vol. 1. IEEE, 2001.
[5] Kessous, Loic, Ginevra Castellano, and George Caridakis. "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis." Journal on Multimodal User Interfaces 3.1 (2010): 33-48.
[6] Deng, Jun, et al. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition." IEEE Signal Processing Letters 21.9 (2014): 1068-1072.
[7] Wang, Kunxia, et al. "Speech emotion recognition using Fourier parameters." IEEE Transactions on Affective Computing 6.1 (2015): 69-75.
[8] Bellegarda, Jerome R. "DATA-DRIVEN ANALYSIS OF EMOTION IN TEXT USING LATENT AFFECTIVE FOLDING AND EMBEDDING." Computational Intelligence 29.3 (2013): 506-526.
[9] Eskimez, Sefik Emre, et al. "Emotion classification: how does an automated system compare to naive human coders?." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.
[10] Phu, Vo Ngoc, Vo Thi Ngoc Chau, and Vo Thi Ngoc Tran. "SVM for English semantic classification

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

in parallel environment." International Journal of Speech Technology(2017): 1-22.
[11] Sharma, Pankaj K., and Prabhat K. Upadhyay. "Cognitive relaying with transceiver hardware impairments under interference constraints." IEEE Communications Letters 20.4 (2016): 820-823.
[12] Yang, Na, IlkerDemirkol, and Wendi Heinzelman. "Cross-layer energy optimization under image quality constraints for wireless image transmissions." Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International. IEEE, 2012.
[13] Chen, Xuemei, and Ruolun Liu. "Multiple pitch estimation based on modified harmonic product spectrum." Proceedings of the 2012 International Conference on Information Technology and Software Engineering. Springer, Berlin, Heidelberg, 2013.
[14] N. Krishnaiah, G. Narsimha,"Web Search Customization Approach Using Redundant Web Usage Data Association and Clustering", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.4, pp.35-42, July 2016. DOI: 10.5815/ijieeb.2016.04.05
[15] Kim, Jeongchan, et al. "An efficient prewhitening scheme for MIMO cognitive radio systems." IEEE Transactions on Vehicular Technology 63.4 (2014): 1934-1939.
[16] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018.
[17]] Danqing Luo, Yuexian Zou, and Dongyan Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," Proc. Interspeech 2018, pp. 152–156, 2018.
[18] N. Krishnaiah, "Automatically Prospecting Feature for Queries from Their Search Impact", International Journal of Engineering and Advanced Technology (IJEAT),ISSN: 2249-8958, Vol 9, Issue 1, October-2019.
[19] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesus´ Villalba, YishayCarmiel, and NajimDehak, "Deep neural networks for emotion recognition combining audio and transcripts," Proc. Interspeech 2018, pp. 247–251, 2018.