A Comparative Study of Clustering Algorithms using RAVDESS dataset

Harshit Jain
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
harshitjaintata@gmail.com

Dhrubo Bhattacharjee
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
dhrubob026@gmail.com

Devansh Tiwari
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
devanshdt2500@gmail.com

Dr.Preksha Pareek
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
preksha.pareek@sitpune.edu.in

Prof.Prachi Kadam
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
prachi.kadam@sitpune.edu.in

***Abstract*—Speech Emotion Recognition (SER) is a hot topic in academia and industry. Although researchers have done a tremendous amount of work in this field, there are still the issues of speech feature choice and the correct application of feature engineering that remains to be solved in the domain of SER. In this research, we have used different clustering techniques like DBSCAN, K-Means clustering, Agglomerative Clustering, Birch Clustering and compare the analysis among them. The results are compared using these clustering techniques using Ryerson Audio-Visual Database of Speech and Song (RAVDESS).**

***Index Terms*—Audio, Emotion, RAVDESS, Dimensionality Reduction , K-Means Clustering, Agglomerative Clustering, Birch Clustering, DBSCAN**

I. Introduction :

Speech is one of the fundamental means by which humans communicate. Because people express themselves through their emotions, it is only natural for them to use emotion in speech to express their feelings clearly.Speech contains both linguistic and nonlinguistic information. A speech signal carries information such as the intended message, the speaker's identify, and the speaker's emotional condition. Effective Communication via language and voice has enabled the exchange of ideas, messages, and perceptions. Acoustic fluctuation and the words themselves are the two most important components in voice-based signalling. Acoustic properties such as pitch, timing, voice quality, and articulation of the speech signal strongly match with the underlying emotion due to the effects of nervous system excitation, increased heart rate, and so on. The fluctuation of these factors serves as the foundation for recognising speech emotions. The procedure for determining a Speech emotion recognition is the detection of a speaker's emotions from their speech output.Finding these emotions shows deeper complexities and aids in issue solving in the present. Emotion recognition from speech is one of the most difficult challenges in the realm of human computer interaction. The development of powerful emotion identification systems is thus advantageous, and the goal of a good emotion recognition system is to be able to Clustering USML Paper 2 replicate human perception in the same way that people can recognise emotions such as anger, sadness, and happiness while conversing.Despite extensive research, there are still a number of challenges to overcome in the field of emotion recognition from speech, such as incomplete databases, poor recording quality, cross-database performance, and difficulties with speaker independent recognition due to each individual's unique speaking style.Speech emotion recognition systems are pattern recognition systems made of three parts: (1) speech signal acquisition, (2) feature extraction, and (3) emotion recognition using classifiers(fig.1). Speech features, qualitative features, spectral features, and Teager energy operator (TEO)-based features are the four categories of features. In this study, we analysed our dataset using various sorts of clustering algorithms.Essentially, we are using clustering to extract comparable acoustic patterns from a given dataset. It is critical to differentiate and extract comparable audio so that the machine can be trained more easily. Audio-based emotion recognition systems can help in mental health assessments. Speech-processing technology can be used to diagnose and assess the severity of illnesses. It can also help with speech therapy, which aims to help persons with speech issues. Furthermore, applications such as health care

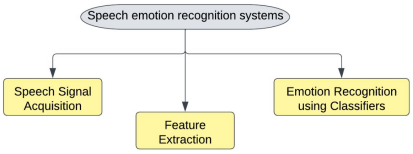

Fig. 1.

Section-3	Proposed Work- 3.1- Dataset Description 3.2- Preprocessing 3.3- Dimensionality Reduction 3.4- Standardization 3.5- Clustering
Section-4	4.1- Result Analysis 4.2-Conclusion 4.3- Future Scope
Section-5	5.1- Reference

TABLE I

and counselling can benefit the most from such automated solutions Speech recognition systems are especially useful where man and machine interaction (MMI) is required, such as in web movies, computer tutorial applications, online learning, call centre communications and so on, because the response in such systems is dependent on the user's sentiment. The purpose of such technologies in contact centre talks would be to detect the caller's urgency and emotional state. This would help to improve the functionality of call centres, particularly those providing health care support to the elderly and emergency call centres. It is also used in vehicle systems to determine the driver's mental state, which is linked to the possibility of reckless driving and subsequent collisions. These devices can help ensure the safety of drivers, passengers, and other road users. It can also be used in forensic and criminal investigations to detect lies. Furthermore, the study demonstrates its use in identifying school violence based on children's speech. Finally, it can improve other AI applications such as marketing, intelligent toys, and music playing that is tailored to the caller's mood.

II. Contribution of this research :

This section defines the contribution of our research concretely. we have implemented different types of Clustering Algorithms. Namely; DBSCAN, K-Means clustering, Agglomerative Clustering, and Birch Clustering, And after all this we are going to use silhouette score to compare the Clustering algorithms using RAVDESS dataset. The rest of the papers is organized as follows:(TABLE 1:)

III. Proposed Work:

A. 3.1 Dataset description -

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a database of emotional speech and song. The RAVDESS dataset contains a total of 7356 files (24.8 GB) of audio and video, spoken and music. It is a multimodal and dynamic set of North American English facial and vocal expressions. The collection features 24 performers who deliver two lexically matched lines in a neutral North American accent. Neutral, calm, happy, sad, angry, terrified, disgusted, and surprised are the eight speech emotions. Each expression has two strength levels: strong and normal, with a neutral expression as an extra. All data is provided in three modes: audio-video, audio solo, and video only. In this investigation, only the audio files, comprising 1440 files (24 actors * 60 trials per actor), are used. Figures 2 and 3 depict the distribution of RAVDESS dataset wave files and filename identifiers, respectively.The audio file naming strategy used in this dataset is depicted in Figure 4

Gender	Count	Trials per Actors	Total Count
Male	12	60	720
Female	12	60	720
Total			1440

Fig. 2. Total Count Of Files

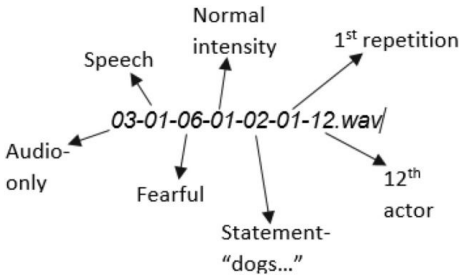

Fig. 3. filename identifier

B. 3.2 Speech Processing -

The speech of the target speaker as well as background noise, the voices of non-target speakers, and reverberation

Modality	01 = full-AV, 02 = video-only, 03 = audio-only
Vocal Channel	01 = speech, 02 = song
Emotion	01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry,06 = fearful, 07 = disgust, 08 =
Intensity	01 = normal, 02 = strong (Note: Strong intensity for neutral emotion is not there)
Statement	01 = "Kids are talking by the door" 02 = "Dogs are sitting by the door"
Repetition	01 = 1st repetition, 02 = 2nd repetition
Actor	01 to 24 Male: Odd numbered actors Female: Even numbered actors

Fig. 4. Naming convention of Audio File

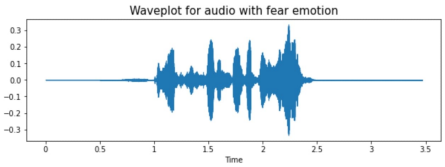

Fig. 5. Raw Audio waveplot

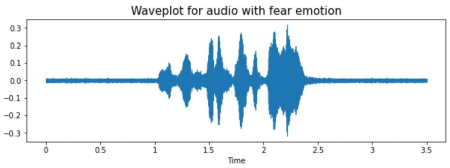

Fig. 6. Noise injected audio waveplot

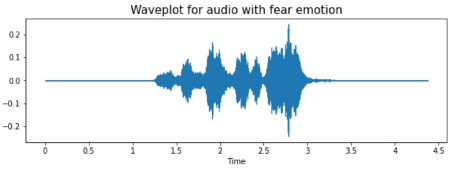

Fig. 8. shifted audio waveplot

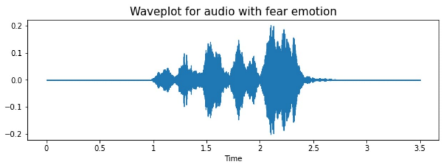

Fig. 9. Changed pitch audio waveplot

are all present in the recoded audio signals. Several speech improvement technologies such as speech processing are used to automatically decrease such interference sounds. Speech processing entails changing or extracting important information from signals by influencing their fundamental properties. The following processes comprise speech processing.

1) 3.2.1 Pre Processing -: The pre-processing step follows data collection. The gathered data would be utilized to train the classifier in an SER system. While only a few of these pre-processing approaches are employed for feature extraction, others take care of feature normalization so that fluctuations in the speech recordings do not affect the feature extraction process

2) 3.2.2 Data augmentation -: We can generate new synthetic data samples by making tiny changes to our existing training set, a process known as data augmentation. To generate syntactic data for audio, we can utilize noise injection, time shifting, pitch and speed variations, and other techniques. Our goal is to increase generalization and make our model insensitive to these perturbations. While librosa (Library for Recognition and Organisation of Speech and Audio) helps with pitch and speed modulation, numpy provides a straightforward method for dealing with noise injection and modifying time. We employed the following augmentation steps: Different augmentation steps that we used were:

1) Noise injection(Fig.6)
2) Trimming the audio(Fig.7)
3) Shifting of audio(Fig.8)
4) Changing the pitch of the audio(Fig.9)

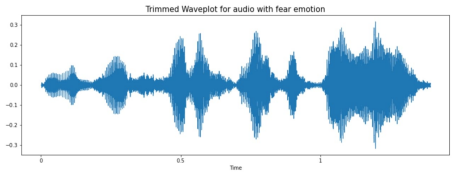

Fig. 7. Trimmed audio waveplot

C. 3.3 Feature Extraction -

The majority of SER's attributes are speech-related. The rate of recognition increases with each well-planned combination of traits that accurately represents every emotion. A variety of attributes have been employed in SER frameworks. There is no commonly agreed feature arrangement for exact and specific classification. The studies that are currently available are all experimental. As we all know, speech conveys information and emotion in varying periods. As a result, depending on our requirements, we can extract either global or local features. There are other extraction methods like:

1) Zero Crossing Rate: The rate of sign changes of the signal during the duration of a particular frame.
2) Energy: The sum of squares of the signal values, normalized by the respective frame length.
3) Entropy of Energy: The entropy of sub-frames normalized energies. It can be interpreted as a measure of abrupt changes.
4) Spectral Centroid: The center of gravity of the spectrum.
5) Spectral Spread: The second central moment of the spectrum.
6) Spectral Entropy: Entropy of the normalized spectral energies for a set of sub-frames.
7) Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8) Spectral Rolloff: The frequency below which 90 percent of the magnitude distribution of the spectrum is concentrated.
9) MFCCs: Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.

However, we concentrated on:

1) 3.3.1 ZCR (Zero-Crossing Rate) - : The Zero-Crossing Rate (ZCR) of an audio frame quantifies how frequently the sign of the signal shifts throughout the frame. In other words, it is calculated by dividing the number of times the signal's value shifts from positive to negative by the duration of the

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} | \operatorname{sgn}[x_i(n)] - \operatorname{sgn}[x_i(n-1)] | ,$$

where $\operatorname{sgn}(\cdot)$ is the sign function, i.e.

$$\operatorname{sgn}[x_i(n)] = \begin{matrix} 1, x_i(n) \geq 0, \\ -1, x_i(n) < 0. \end{matrix}$$

Fig. 10. Zero-Crossing Rate Equation

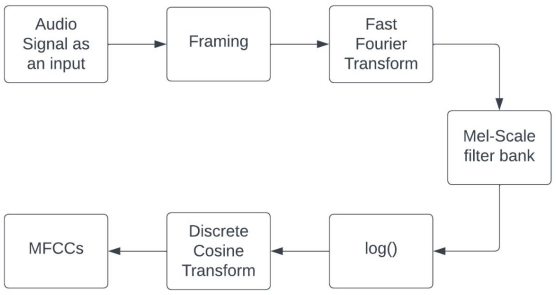

Fig. 11. MFCC Flow chart

frame.(Fig. 10)

2) 3.3.2 MFCC (Mel-Frequency Cepstral Coefficients) - : The Mel Frequency Cepstral Coefficient is the most often used spectral characteristic in automatic speech recognition(MFCC). MFCC seeks to recreate how our ears work by processing speech waves linearly at low frequencies and logarithmically at high frequencies. The envelope of the short-time power spectrum, which characterizes the geometry of the vocal tract, is represented by MFCCs. During framing, the characteristics of the voice signal are divided into frames for proper analysis. Because speech signals are not stationary, their properties vary every 15 milliseconds. To generate MFCC, the short-time discrete Fourier transform is employed to separate the utterances into distinct segments before translating them into the frequency domain. The Mel filter bank is used to determine many sub-band energies. The logarithm of the relevant subbands is then computed. Finally, the inverse Fourier transform is employed to compute MFCC.
Mel(f) = 2595
log10 1+f/1000, where f represents the actual frequency of speech.
The link between speech frequency and Mel scale can be calculated as follows:
Frequency (Mel) = [2695 log (1 + f(Hz))/700].
(Fig. 11)

3) 3.3.3 Mel Spectrogram - : We use the Mel spectrogram to provide our models with sound data that is close to what a human would hear. Filter banks are applied to raw audio waveforms to generate the Mel spectrogram. Following this technique, each sample has a 128 x 128 shape, indicating the

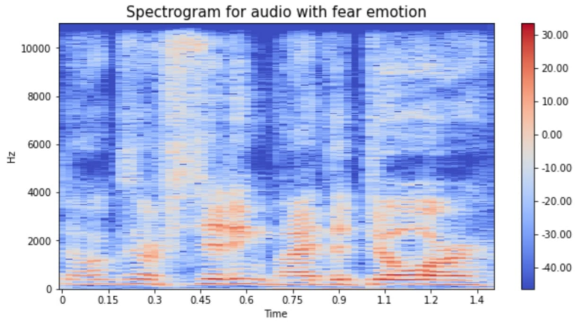

Fig. 12. Mel Spectogram

use of 128 bits filter banks and 128 time steps per clip.An Example of Mel Spectrogram is shown in (Fig. 12): From all these, MFCC fits best for feature extraction.

D. 3.4 Standardization -

It's conceivable that you will run over datasets with bunches of mathematical commotion worked in, like change or in an unexpected way scaled information, so a decent preprocessing is an unquestionable requirement before contemplating AI. A decent preprocessing answer for this sort of issue is frequently alluded to as standardization. Standardization is a preprocessing strategy used to change constant information to make it look typically dispersed. In scikit-learn this is much of the time an essential step in light of the fact that many models expect that the information you are preparing on is regularly disseminated, and in the event that it isn't, your gamble biasing your model.You can standardize your information in various ways, scaling is one of them.

1) 3.4.1 Standard Scaler - : Standard scaling is an standardization process commonly used in machine learning for data prepocessing and statistical analysis. It's purpose is to transform numerical data so it has mean of zero and a standard deviation of one. This process involves subtracting the mean of the data from each data point and then divide it by standard deviation. This transformation centers the data at zero and scales it to have unit variance.After this scaling process, the data would have mean of 0 and SD of 1.

E. 3.5 Dimensionality Reduction -

The process of transforming data from a high-dimensional space into a low-dimensional space with the goal of keeping the low-dimensional representation as close as possible to the inherent dimension of the original data is known as dimension reduction. Working with high-dimensional spaces can be undesirable for a variety of reasons, including the fact that the data analysis is typically computationally intractable and that the raw data are frequently sparse as a result of the curse of dimensionality (hard to control or deal with). Dimensionality reduction is frequent in disciplines like signal processing and voice recognition that deal with high numbers of observations and/or large numbers of variables. Data downsizing is an

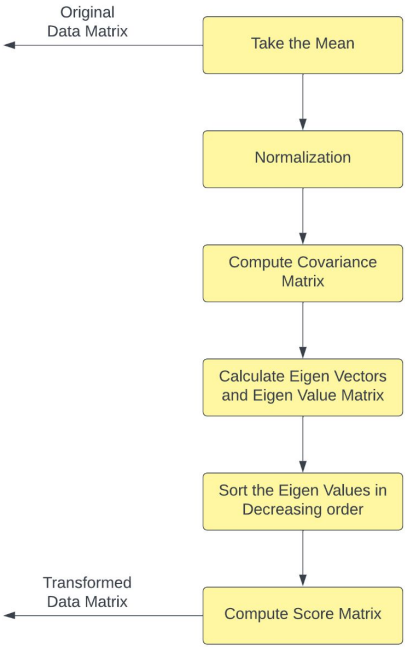

Fig. 13. PCA Algorithm

essential strategy. The dimensionality of data can be reduced using a variety of approaches:

1) Principal Component Analysis (PCA)
2) Non-negative matrix factorization (NMF)
3) Linear discriminant analysis (LDA)
4) Random Forests

We only used PCA:

1) 3.5.1 PCA (principle component analysis) - : When it comes to SER, PCA appears to be one of the most widely used approaches. Principle component analysis (PCA) is a preprocessing linear transformation approach that identifies a subspace whose basis vectors match the highest variance in the original space. Typically, PCA looks for the surface with the lowest dimensionality onto which to project the highdimensional data. PCA functions by taking into account each attribute's variance since a high attribute demonstrates a solid split between classes, which lowers the dimensionality. (Fig. 13) The steps for performing PCA are:

1) Standardize the data.
2) : Compute the covariance matrix.
3) Compute the eigenvectors and eigenvalues of the covariance matrix.
4) : Choose the number of principle components to retain. This can be done by examining the scree plot or by selecting a threshold for the amount of variance explained.
5) Transform the data into the new coordinate system defined by the principle components. This is done by multiplying the original data matrix X by the matrix of eigenvectors V, selecting the desired number of principal components k.
6) Interpret the results.

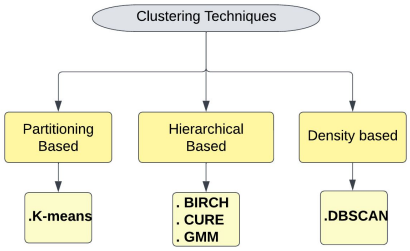

Fig. 14. Clustering Techniques

F. 3.6 Clustering Algorithms -

Clustering or cluster analysis is an AI procedure, which groups the unlabelled dataset. It tends to be characterized as "An approach to gathering the data of interest into various clusters, comprising of comparable data of interest. The items with the potential similitudes stay in a gathering that has less or no likenesses with another gathering."It does it by discovering a few comparative examples in the unlabelled dataset like shape, size, variety, conduct, and so on, and isolates them according to the presence and nonattendance of those comparable examples.It is an unaided learning strategy, subsequently no management is given to the calculation, and it manages the unlabeled dataset.Subsequent to applying this clustering strategy, each cluster or gathering is given a group ID. ML framework can utilize this id to improve on the handling of huge and complex datasets.
Taxonomy of Clustering Algorithms: There exist many measures and initial conditions which are responsible for numerous categories of clustering algorithms A widely accepted classification frames clustering techniques(Fig. 14).

1) 3.6.1 Partitioning BASED - :

1) K-Means CLUSTERING

K-means is an undeniably popular partitioning algorithm. Many specialists from other domains have discovered, rediscovered, and examined it, including Steinhaus (1965), Ball and Hall (1965), Lloyd (suggested 1957 - published 1982), and Macqueen (1967). It is distance-based, and data is partitioned into pre-determined groups or clusters by definition. Euclidean or cosine distance metrics could be utilized. Initially, a fixed K cluster centroids are marked at random; k-means reassigns all points to their nearest centroids and recomputes centroids of newly formed groups. This process is repeated until the squared error converges.The following steps summarise the k-means function:
1. Initialize a K partition based on previous information. A cluster prototype matrix A=[ai.....aj] is created. Where a1,a2,a3 ... are cluster centers. Data set D is also initialized.
2. In the next step assignment of each data point in the dataset(di) to its nearest cluster(ai) is performed.

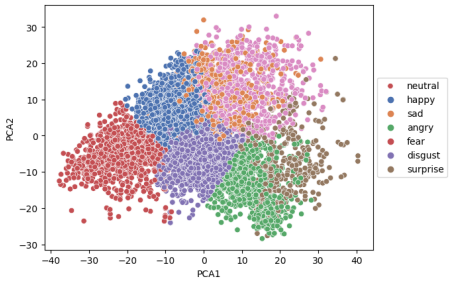

Fig. 15. K-Means 2D plot

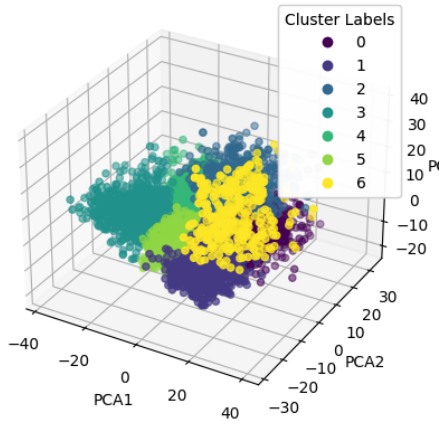

Fig. 16. K-Means 3D plot

3. Cluster matrix can be recalculated considering the current updated partition or until ai,aj,ak..... Show no further change.
4. Repeat 2 and 3 until convergence has been reached.
K-means is perhaps the most well studied method, which is why there are so many variations and improved versions of it, yet it can exhibit considerable sensitivity to noise and outliers in data sets. Even though a point is far from the cluster centroid, it can still be driven to the centre, resulting in a deformed cluster shape. In the beginning, K- means does not clearly describe a universal technique of determining total number of partitions; instead, this approach relies largely on the user to provide the number of k clusters in advance. Furthermore, k-means cannot be used with categorical data. Because k-means assumes that the user will provide the initial assignments, it can produce repeated results with each iteration.The k-means can addresses this problem by attempting to choose better starting clusters. Different output of the K-means Clustering are given in the figure (Fig. 15),(Fig. 16).

2) 3.6.2 Hierarchical CLUSTERING - : This category is a cluster analysis paradigm that generates a sequence of nested partitions (clusters) that can be visualised as a tree or a hierarchy of clusters known as a cluster dendrogram. Hierarchical trees can show data at several levels of abstraction. When represented as a tree, this hierarchy can have the lowest level, say, leaves, and the highest level, say, roots. Each point in the leaf node has its own cluster, whereas the root node has all points in one cluster. The dendrogram can be sliced at intermediate levels to yield clustering findings; relevant clusters can be found at one of these intermediate levels.

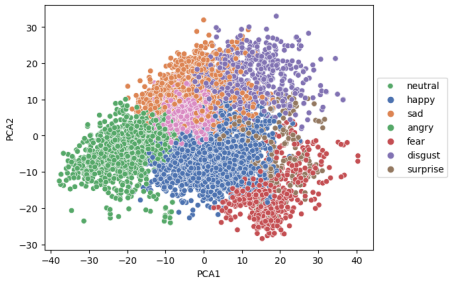

Fig. 17. BIRCH 2D plot

1) BIRCH CLUSTERING: Birch's "Balanced iterative reducing and clustering using hierarchies" In clustering analysis, the Birch solves two major concerns:
1. Scalability when dealing with large amounts of data.
2. Robustness against noise and outliers.
Birch offers a novel data structure, the (CF) tree clustering feature, to achieve the aforementioned goals. The CF tree can be thought of as a tuple that summarises the information kept about a cluster. Instead of using the data directly, the CF tree can compress it and generate a large number of tiny points or nodes.These nodes function as miniature clusters and display a summary of the original data. This is a height balanced CF tree with two parameters: branching factors B and T, which is the threshold. Every internal vertex in the CF tree is made up of entries described as [CFi,childi], I = 1....k, where CFi is a summary of the cluster i and is defined as a tuple. CF i = (Ni,LS,SS), where Ni is the number of data objects in the cluster. LS = the linear sum of the N data points, SS = the object squared sum. The CFs are saved as leaf nodes, whilst non leaf nodes are made up of the sum of all the CFs of their children. When an object is inserted into the closest leaf entry, the two parameters B and T control the maximum number of children per non leaf node and the maximum diameter of sub clusters stored in the leaf node, respectively. Birch builds a framework that may be kept in main memory in this fashion. After generating a CF tree, the next step is to do clustering using an agglomeration hierarchical or any of the sum of square error based algorithms.It is quite fast and with multiple scans it gives improved results but again the inability to deal with non-spherical shaped clusters stands as the biggest drawback of birch.Arbitrary shaped clusters cannot be identified by birch as it uses the principal of measuring diameter of all clusters for determining their boundary. Different output of the Birch Clustering are given in the figure: (Fig. 17),(Fig. 18).
2) CURE :It was created to recognise more complicated cluster structures . It is more resilient to outliers. Cure is

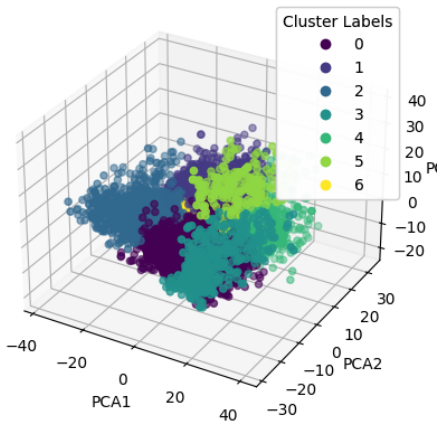

Fig. 18. BIRCH 3D plot

a process of agglomeration. Instead of a single centroid, it assumes multiple distinct fixed points as clusters and use a fragment of m to shrink these disparate locations towards centroids. At each iteration, these scattered dots represent the cluster, and the pair of clusters with the closest representatives is fused together. This feature allows CURE to correctly recognise clusters and makes it sensitive to outliers. This algorithm incorporates two improvements:

1. random sampling
2. partitioning

These enhancements contribute to the cure's scalability. Random sampling can have an impact on memory requirements, making it a costly option for data bases.

3) GMM: Gaussian Mixture Model (GMM) clustering is a form of unsupervised machine learning algorithm used to group or cluster data points. Data points are modelled as a combination of many Gaussian distributions in this approach, with each Gaussian distribution representing a different cluster. The GMM algorithm seeks to determine the Gaussian distribution parameters (i.e., mean and variance) that best describe the underlying distribution of the data. GMM is a probabilistic model, which means it computes the probability of each data point belonging to each cluster based on Gaussian distribution parameters. The technique iteratively adjusts the Gaussian distribution parameters to maximise the likelihood of the observed data given the model. In practise, GMM clustering is frequently used for image segmentation, anomaly detection, and natural language processing. It's also utilised in more advancedmachine learning methods like Gaussian mixture regression and Hidden Markov Models.

3) 3.6.3 Density Based Clustering - : Martin Ester, Hans-Peter Krigel, JorgSander, and Xiawoei suggested it in 1996.one of the most widely used density-based algorithms. If an object belongs to a cluster, the density in its vicinity must be sufficiently high[16]. This set of core objects with overlapping neighbourhoods is used to build a cluster skeleton. The boundary of clusters is represented by points within the neighbourhood of core objects, whereas the remainder is essentially noise. It necessitates the use of two parameters.
1) is the beginning point
2) Minpts is the smallest number of points needed to make a dense zone.
The following steps can elaborate the algorithm further:
1 An unvisited random point is usually taken as the initial point.
2. A parameter E is used for determining the neighborhood (data space)
3. If there exist sufficient data points or neighborhood around the initial random point then algorithm can proceed and this particular data point is labeled as visited or else the point is labeled as a flaw in data or outlier.
4. If this point is considered a part of the cluster then its E neighborhood is also the part of the cluster and step 2 is repeated for all E. this is repeated until all points in the cluster are determined.
5. Another initial data point is processed and above steps are restated until all clusters and noise are discovered.
Although this technique performs well against noise, it can fail when evaluated in high-dimensional data sets and is sensitive to Minpts.In terms of producing clustering of various shapes, this algorithm outperforms k-means.

4) 3.6.4 Silhouette Score - : The silhouette score is a metric used to assess the output quality of a clustering method. It compares an object's similarity to its own cluster to that of other clusters. A silhouette score runs from -1 to 1, with 1 indicating that an object is very similar to its own cluster and -1 suggesting that it is very distinct to other clusters. A score of -1 indicates the inverse, whereas a score of 0 shows that the object is equally similar to its own cluster as well as other clusters.To calculate the silhouette score, the following steps are taken for each object in the dataset:
1. Compute the average distance between the object and all other objects in its cluster. This is called the intra-cluster distance (a).
2. Compute the average distance between the object and all objects in the nearest cluster. This is called the nearest-cluster distance (b).
3. Calculate the silhouette score for the object as (b - a) / max(a, b).
4. Compute the average silhouette score for all objects in the dataset.
A high silhouette score suggests that the grouping is suitable since the objects are well-clustered and distinct from those in other clusters. A low silhouette score suggests that the clustering is ineffective since the objects are poorly clustered and comparable to objects in other clusters.

IV. Analysis :

A. 4.1 Result Analysis -

With using any dimensionality reduction approach, PCA the Silhouette score we obtained for DBSCAN Clustering was -

Algorithm	Silhouette Score
DBSCAN	-0.1713
K-Means	0.2276
Agglomerative Clustering	0.2194
BIRCH Clustering	0.1569

TABLE II

0.1713, whereas the Silhouette score we obtained for K-means was 0.2276, the Silhouette score we obtained for Agglomerative Clustering was 0.2194 and for BIRCH Clustering it was 0.1569. So according to the score K-Means performed well among all the others and can be used for the provided dataset.(TABLE 2.)

B. 4.2 Conclusion -

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is a collection of audio and video recordings of actors performing emotional speech and song. Depending on the study question or application, clustering on this dataset can be valuable for a variety of reasons. Here are a few reasons why clustering could be beneficial for the RAVDESS dataset:
1) Identifying patterns: Clustering can aid in the identification of patterns in emotional speech and song recordings, such as similarities or differences in the performances of the actors or similarities or differences between emotional categories. Researchers will be able to better grasp how emotional expression is conveyed in speech and song as a result of this.
2) Data visualisation: Clustering can also be used to visualise data by emphasising patterns or contrasts. A clustering algorithm, for example, may group together recordings with similar emotional expressions, which can then be shown as a heat map or scatter plot.
3)Extraction of features from audio or video recordings: Clustering can also be used to extract features from audio or video recordings. A clustering algorithm, for example, may find common audio or visual elements linked with specific emotional expressions. This can help in the development of automated emotion recognition systems.
In conclusion, clustering on the RAVDESS dataset can assist academics in better understanding emotional expression in voice and song, visualising patterns in data, and extracting features for use in automated emotion identification systems

C. 4.3 Future Scope -

To the best of our knowledge, this is the first time we have used the RAVDESS dataset to apply dimension reduction and clustering approaches to our model. We would investigate further dimensionality reduction methods like convolutional autoencoders and denoising encoders and do clustering on basis of that for future work.

V.

[2] [1] [3]

REFERENCES

[1] Sofia Kanwal and Sohail Asghar. Speech emotion recognition using clustering based ga-optimized feature set. *IEEE Access*, 9:125830–125842, 2021.

[2] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875, 2020.

[3] Tshephisho Joseph Sefara. The effects of normalisation methods on speech emotion recognition. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8, 2019.