Volume 65, Issue 5, 2021

Journal of Scientific Research

of

The Banaras Hindu University

Mutual Information Based Efficient Model for Speech Emotion Recognition

Shivangi Srivastav*[1], Rajiv Ranjan Tewari[2]

*[1]Center Of Computer Education, University Of Allahabad, Prayagraj, meshiwangi6nov@gmail.com
[2]Department Of Electronic and Communication, University Of Allahabad, Prayagraj, tewari.rr@gmail.com

Abstract: **Speech is an important characteristic for identifying a person from human interaction/communication in everyday life. Automatic Speaker Recognition (ASR) is the method to detect individuals by observed highlights of voice signals. Speech signals are excellent mediums for communication which continuously transmit rich, valuable information, such as the feelings of a speaker, sexual orientation, complementarity and other fascinating characteristics. The fundamental goal in identification of speaker is to remove valuable highlights and enable important instances of models of speaker. Hypothetical description, complete sensation structure and methods of communication of the emotion are included. In view of the various classifiers and the varied methods for extracting highlights, a SER framework was created for this research. In this study, different machine learning methods are examined to find decision limits in the audio signal space. Furthermore, this technique is new in enhancing the performance of conventional machine learning algorithms by selecting features based on information theory. The better accuracy is 96 percent utilising a random forest algorithm with the technique of selection of common information.**

Index Terms: **Speech recognition, feature extraction, SVM, linear multivariate regression, spectral modulation characteristics, machine learning.**

I. Introduction

A key focal point of the language system is speech recognition. Recognition of speech is a perplexing order errand grouped by various mathematical processes: acoustic phonetic technique, approach to design recognition, approach to man-made consciousness (artificial neural networks), dynamic dime distortion, close connectionist draws (artificial neural network), and vector machine help.

Hidden Markov models are good for the ambiguous or insufficient knowledge that results from confounding sounds, word game plan assortments, homophone words, speaker rate and function, and clear speech recognition effects. Nevertheless, there are a variety of HMM obstacles: thickness components of the model are inadequate to reflect the data structure; raising the specialist intensity of the model is not the best measure to achieve ideal execution of talk affirmation; difficult to evaluate the mix-ups of an HMM device seeking to enhance its display [2, 3].

In spite of the way the speakers are not really present, these excellent properties allow researchers to perceive between speakers when bringing are organized over telephones. Machines may settle with the outflows of speakers, similar to individuals, through such characteristics. Speaker articulations are set up from the assembled dataset with the ML algorithm, and the test articulations differentiate speakers a while later.

One of the review territories in sound handling and example recognition is the speaker acknowledgment. Speaker acknowledgment integrates ID, confirmation (validation), order, and likewise, division, following and so forth [1].

Proposed work, the device purpose to identify the speaker from a collection of recognized speakers, such as evaluating the speech of the test speaker against all usable speaker models and returning the speaker ID generated by model with the nearest coordinate and no guarantee of character. It is defined as a shut-down set situation at the stage where the sound of the test speaker is chosen from the recognised set. Similarly, the test speaker can be from beyond the predefined meeting and it becomes an open-set case. The material ward and text free are the two modalities for the identification of programmed speakers. Text of speech used for planning and checking should be the same in content ward speaker recognition text where there is no compulsion on the content in content autonomous. The book free shut set speaker identifiable proof method is revised in this job execution evaluation.

The square chart of the identifiable proof scheme for the speaker appears in Fig. 1Manuscript Organization.

* Corresponding Author

DOI: 10.37398/JSR.2021.650537

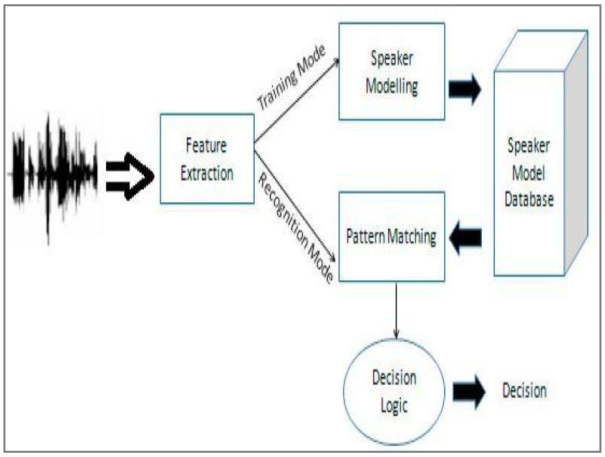

Fig. 1. Speech identification system

II. LITERATURE REVIEW

The field of AI, combined with cognitive science, is evolving rapidly. As of late, to validate singular distinguishing facts, voice biometrics have been used.

Because of its highlights of effortlessness, individuality and all-inclusiveness, the human voice is the most helpful vehicle of communication. The benefits of speaker ID in comparison to other biometric control systems are as follows:

1) Voice is efficiently available, easy to use and costs are minimal.

2) Voice is anything but difficult to get and equally easier for customers to view people.

As speech acknowledgment system need to operate under a broad assortment of circumstances, subsequently, such system should be vigorous to extraneous varieties incited by various acoustic influences, for example, transmitting channel, speaker contrasts and foundation commotion. A significant portion of the speech applications perform advanced channels to enhance order execution, where the optimal evaluation of expression is sought by going through a straight channel of uproarious expression.

In all inclusive practice, the human voice is used to exchange with each other, information. Recognition of the speaker alludes to the identifiable evidence of speakers relying on the vocal characteristics of the human voice. Due to its wide variety of applications, for example legal voice search to identify offenders by government law requirement agencies, speaker recognition has become a region of severe review [1], [2]. Extraction of features in the measure of speaker acknowledgement assumes a key job because it fully affects the presentation of a model of speaker acknowledgement order. As of late, numerous analysts have suggested novel highlights in the field of speaker recognition that have been shown to be helpful in arranging human voices viably. Isolated the remaining stage and MFCC highlights from the NIST 2003 dataset's 149 male speaker expressions for framing a master function vector. As a contribution to an auto-affiliated neural organisation classifier, the designers took care of the extricated MFCC function and obtained approximately 90 percent classification precision. However, for complex datasets, such as Libri Expression, the suggested highlights and classification may not be appropriate. In order to organise speakers using different time-area factual highlights and ML classifiers, W.-C. Chen et al. [11] carried out a similar report; by using the multilayer perceptron classifier, they acquired the highest accuracy of about 94 percent. Given the fact that the exploratory aftereffects of the examination accomplished great characterization exactness, the outcomes can't be summed up to a more comprehensive scale in light of the fact that the creators used only 16 speaker voices.

To coordinate speakers using a SVM calculation, the analysis merged deep learning-based and MFCC highlights. 92 percent arrangement accuracy was reached via the exploratory results. The findings are however, promising. Be that as it may, there are a few shortcomings in the dataset used in the assessments. Just 10 speaker phrases were used in the tests to begin with. Second, only a single word was used in any phrase. In this way, for diverse human voices, the combination-based highlights suggested by the writers could be wasteful and inadequate.

Clustering-based MFCC highlights combined with an ANN classifier were explored by D. B. A. Mezghani, et al.[12] to sort 22 speakers from the ELDSR dataset. 93 percent characterization accuracy was obtained by the exploratory after-effects of the investigation. 86 percent and 88 percent grouping accuracy is shown by the consequences of the analysis that used the suggested highlight extraction strategies. Mirsamadi, et al.[14] suggested a collection of discriminative highlights from the MEPCO speech dataset to order 50 speaker phrases. RASTA-MFCC highlights have been isolated by these developers to identify speaker expressions. To get acquainted with the classification rules, the deleted highlights were entered into a GMM-general foundation model classifier. 97 percent classification accuracy was reached in the results. Given the fact that the findings demonstrated fair accuracy of classification, in view of the fact that the test only used six terms, with one expression going on for only 3s, they cannot be extended to a more comprehensive scale. In this way, RASTA-MFCC highlights can end up being immaterial for speaker expressions that are over 3s long.

The current review proposed new MFFCT highlights to order speaker expressions in order to enhance the characterization accuracy of the LibriSpeech dataset for speaker recognition. In addition, a DNN to create a speaker ID model was applied to the separate MFCCT highlights. In the resulting fields, the subtleties of the proposed highlights and model are spoken of areas.

III. FEATURE EXTRACTION

Feature extraction is the mechanism to observe a series of highlights for each brief timeframe enclosure of the data signal, assuming that quite a limited portion of speech is sufficiently fixed to enable substantial modelling. Highlights in speech handling are sorted into;

1. Spectral short-term
2. Source of speech
3. Temporal-spectro
4. Prosodic
5. Based on their physical interpretations, high-level characteristics

There are countless limits in the discourse signal, which represent the energetic features. What characteristics should be used is one of the staying centers in inclination confirmation. Different critical characteristics are isolated in late reception. The most widely used representation of the ridiculous feature of voice signals is the Mel-recurrence cepstrum coefficient (MFCC). This are the best for appreciation of dialogue as it takes into account the affectability of human acumen with respect to frequencies.

The feature vector of each MFCC is 60-dimensional. Otherworldly Balance Highlights (MSFs) are isolated from a long-haul spectro-transient representation powered by hearing. These characteristics are accomplished by imitating the spectro-worldly (ST) handling in the human hear-able environment and by jointly considering the normal acoustic frequency with modulation frequency. The speech signal is first disintegrated by a hear-able filterbank to obtain the ST representation (19 channels altogether). To frame the balance signals, the Hilbert envelopes of the fundamental band yields are figured. Additionally, a modulation philtre bank is added to the Hilbert envelopes to perform recurrence checking. The tweak signals are referred to as modification spectra in the otherworldly material, and the proposed highlights are subsequently referred to as regulatory ghost highlights (MSFs)[5]. In conclusion, the ST representation is framed as a part of normal acoustic recurrence and modulation frequency by estimating the energy of the decayed envelope signals. A part is given by the energy, presumed control over all casings in each spectral band. A hear-able philtre bank with N 1⁄4 19 channels and a modulation philtre bank with M 1⁄4 5 channels are used in our research. In this job overall, 95 MSFs are calculated from the ST representation.

The objective of the selection of features in ML is to "decrease the quantity of features used to describe a dataset in order to improve a learning algorithm's exhibition on a given undertaking." Selection of features (FS) means selecting a subset of essential features from the first ones, as indicated by a certain basis for significant evaluation, which usually prompts greater accuracy of recognition[18]. It will certainly reduce the learning algorithm's running season. We present a good feature selection method used in our work in this section, a recursive feature element ending with linear regression (LR-RFE).

The most un-significant highlights are removed from the present arrangement of highlights at that stage. On the pruned set, the system is recursively rehashed until the ideal number of characteristics to choose is ultimately reached. In this work, we introduced the technique of recursive disposal of features to include positioning through the ranking of features through the use of simple linear regression (LR-RFE)[32]. In addition, another analysis uses RFE with another straight model, such as SVM-RFE, which is a measurement of SVM-based part determination made by [15]. Using SVM-RFE, main and important capabilities were selected by Guyon et al. Despite enhancing the order accuracy rate, it can reduce computational time for grouping.

IV. CLASSIFICATION METHODS

For discrete emotion order, multiple ML algorithms were used. The purpose of this algorithm is learn from the examples of the preparation or then use this to find out how to order new perceptions. There is therefore no definitive answer to the learning algorithm's decision; each procedure has its own desirable circumstances and impediments. Hence, we agreed to look at the presentation of three distinctive classifiers here.

Illustration 4. RNN and computational time unfolding that is involved in its forward computation[10].

The LRC algorithm represented as Algorithm 1 was somewhat altered[39]. In contrast to various classifiers, it can have an outstanding grouping execution, especially for limited knowledge preparation [11]. You will find the SVM hypothetical base in [30]. In [31], a MATLAB stash updating SVM tool is accessible uninhibitedly. In this work, a polynomial bit is examined.

Compared with Decision Trees and Random Forest, support vector machines are suitable classifiers. It can control information spaces that are nonlinear. SVM is suitable for continuous speech recognition in medium to wide jargon. Using SVM, acoustic states of speech during preparation and testing can be characterized.

V. ALGORITHM

A. *Texture Features*

A growth test has to be determined at both apparent and surface levels. Tonal feature governs with minor force variations in a small district. However, textural characteristics are used for unexpected power fluctuations (hazardous behavior) to evaluate its severity. There are two important areas to consider in such an examination, first the fixed size and second the course of the test. GLCM analyses the spatial connection in order to define low co-event estimates. It counts the times that I coexist with the dark j, while gazing a certain manner. This piece explores all eight possible neighborhood rooms. By normalizing all eight directional GLCM-metrics the 2-D probability matrix Ᵽ (j,k) is

produced. Here I, jµl (inconceivable dark characteristics in a photo) and Ᵽ(j,k) are the LXL square matrices. We obtain Haralick textural features[34] for highlight numbering represented by Table I.
Where µx, µy signify minimal methods given by,

Algorithm 1. Texture Features

```
1. $\mu x = \Sigma\Sigma j * \text{Ᵽ}(j, k)$ and $\mu y = \Sigma\Sigma j * \text{Ᵽ}(j, k)$
2. σx, σy represents marginal variance calculated as,
   $\sigma x = \Sigma\Sigma(j - \mu x)2 * \text{Ᵽ}(j, k)$
   $\sigma y = \Sigma\Sigma(k - \mu y)2 * \text{Ᵽ}(j, k)$
3. $\text{Ᵽ}_{mean}$ denotes the avg of matrix Ᵽ.
4. $\text{Ᵽ}_{x+y}(j)$ represent sums of probabilities; where $2 \leq j < 2l$.
5. $\text{Ᵽ}_{x-y}(j)$ represents differences of probabilities; where $0 \leq j <$ l-1.
6. For calculating complete entropy of signal, count $S_{XY}$ by,
   $S_{XY} = -\Sigma\Sigma\ \text{Ᵽ}(j, k) * \log(\text{Ᵽ}(j, k))$
7. Correspondingly, $S_X$ and $S_Y$ denotes entropies related with marginal probabilities of $\text{Ᵽ}_X$ and $\text{Ᵽ}_Y$.
8. And, $S_{XY}1$ and $S_{XY}2$ is given by,
   $S_{XY}1 = -\Sigma\Sigma\ \text{Ᵽ}(j, k) * \log\{\text{Ᵽ}_X(j) * \text{Ᵽ}_Y(k)\}$
   $S_{XY}2 = -\Sigma\Sigma\ \text{Ᵽ}_X(j) * \text{Ᵽ}_Y(k) * \log\{\text{Ᵽ}_X(j) * \text{Ᵽ}_Y(k)\}$
```

Table-1: GLCM Based Haralick Texture Features

Texture features	**Equation**
Auto correlation	$\sum\sum j*k*\text{Ᵽ}(j,k)$
Contrast	$\sum\sum (j-k)^2*\text{Ᵽ}(j,k)$
Correlation1	$\frac{\sum\sum j*k*\text{Ᵽ}(j,k)-\mu_x\mu_y}{\sigma_x\sigma_y}$
Correaltion2	$\sum\sum\frac{((j)\mu x)*((k)-\mu y)*\text{Ᵽ}(j,k))}{\sigma_x\sigma_y}$
Cluster shade	$\sum\sum(j+k=\mu y-\mu x)3*\text{Ᵽ}(j,k)$
Cluster prominence	$\sum\sum(j+k=\mu y-\mu x)4*\text{Ᵽ}(j,k)$
Dissimilarity	$\sum\sum \text{abs}(j-k)*\text{Ᵽ}(j,k)$
Energy	$\sum\sum\text{Ᵽ}(j,k)2$
Entropy	$\sum\sum\text{Ᵽ}(j,k)*\log\{\text{Ᵽ}(j,k)\}$
Homogeneity1	$\frac{1}{1+(j-k)2}*\text{Ᵽ}(j,k)$
Maximum probility	$\text{MAX}j,k\ \text{Ᵽ}(j,k)$
Sum of squares	$\sum\sum\text{Ᵽ}(j,k)*(j-\text{Ᵽ}_{mean})2$
Sum average	$\sum 2l_j=1\ j=1\ (i+1)*\text{Ᵽ}_{x+y}(j)$
Difference variance	$\sum l-1\ j=0j2*\text{Ᵽ}_{x-y}(j)$
*Sum entropy	$\sum 2l-1\ j=1\ \text{Ᵽ}_{x+y}(j)*\log\{\text{Ᵽ}_{x+y}(j)\}$
{Sum variance	$\sum 2l-1\ I=1\{(j+1)-f1\}2*\text{Ᵽ}_{x+y}(j)\}$
Difference entropy	$\sum l-1\ I=0\ \text{Ᵽ}_{x+y}(j)*\log(\text{Ᵽ}_{x-y}(j))$
Information measure of	$\frac{S_{XY}-S_{XY}2}{\max(S_X,S_Y)}$

Table-2: List of features present in an audio signal

Feature Name	**Description**
Zero Crossing Rate	"The rate at which the signal changes its sign."
Entropy of Energy	"The rate at which the signal changes its sign."
Entropy of Energy	"The value of the change in energy."
Spectral Centroid	"The Spectrum middle value."
Spectral Spread	"The value of the bandwidth in the spectrum."
Spectral Entropy	"The value of the bandwidth in the spectrum."
Spectral Rolloff	"The value of the frequency under which 90% of the spectral distribution occurs."
MFCCs	"Mel Frequency Cepstral Coefficient values of the frequency bands distributed in the Mel-scale."
Chroma Vector	"Every pitch class energy's 12 values."
Chroma Deviation	"The Chroma vectors quality deviation value."

B. Feature Selection

The high dimensionality of features is a revile for characterization issues. Countless unessential highlights may help calculation load, over-fitting and complex interpretability of came about model. Hence, we need to choose a bunch of significant features from crude component data set. It tends to be performed utilizing features based separating techniques or Feature Selection (FS) systems. During classification of examples M related with marks N, there exists an upset common data that can be depleted to improve the classifier execution. Bayes error[35] of anticipating class N from design M is lower limited by Fano's imbalance and upper limited by half of contingent entropy as given by below equation.

$$\frac{H(N)-I(M;N)-1}{\log(|N|)} \leq P(g(M) \neq N) \leq \frac{1}{2}H(\frac{N}{M})$$

Table 3 Different Feature Selection Methods

Feature selection	**Method**	**Mathematical formulation selection criterion**
Maximum relevancy mimimum redundancy (MRMR)	J mrmr(Mt)=max(ixt;N)- $\frac{\sum_{k-1}^{n-1}\frac{IM,Nk}{n-1}}{n-1}$	Concentrate on the minimul redundancy of chosen feature Mt.
Joint mutual information (JMI)	J jmi (Mt)=max($\sum_{k=1}^{n-1} I(Mt\ Mk;N)$)	Includes set of distinctly correlated points amongst the complete set of n points to desrease the classification error.
Conditional mutual information maximization (CMIM)	JCMIM(x t)=max (min $\forall kI(Mt:N/Mk)$)	The chosen function Mt of specific type N need to now not engage with already chosen characteristic Mk
Double input symmetric relevance (DISR)	J $_{\text{DISR}}$(Mt)=max ($\sum_{k-1}^{n=1}(\frac{Ix\ tMk:N}{H(MtMk:N}$)	Uses joint entropy H to normalize JMI for compensating the bias in excessive arity elements
RELIEF	Jmrmr(Mt)=W i-(mi-near-hit)2+	Iteratively assign weights to aspects the use of coefficient WiA RANDOM occasion xi is chosen to decide its distance to instances of nearest hit

VI. RESULTS & ANALYSIS

The testing was conducted on a 5.5GHz Intel Core I5 CPU during MATLAB2013. Windows 7 is a key step in the execution of MATLAB instructions from higher to lower levels. In this study we examined the presentation of machine learning algorithms on the characteristics derived from EMO-DB Berlin emotional speech database[14] speech sound documents.

Seven kinds of emotional classes are essentially included in the study, relating to happiness, anger, boringness, disgust, anxiety, neutrality and sorrow of people spread across 526 German-language voice files.

The functional matrix is evaluated for JMI, MRMR, CMIM, fisher, relief and MIM feature selection methods in order to achieve a newly generated feature matrix that is identical to the original one. However, the features are classified according to their mutual information with the next rank and class label. The whole dataset is separated into a size 421 and 105 training and test data set. During the simulation an increase in features is sliced and used as input for machine learning algorithms for the development and definition of multiclass borders in feature areas, which will later be evaluated over a test data set to get the accuracy of the test classification. The results of the simulation for different machine learning methods utilizing training and test data are shown in figures 1 to 7 correspondingly.

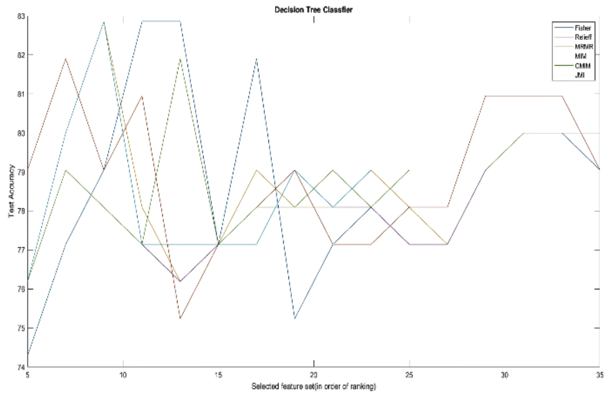

Fig.2. Training Accuracy Using Decision Tree

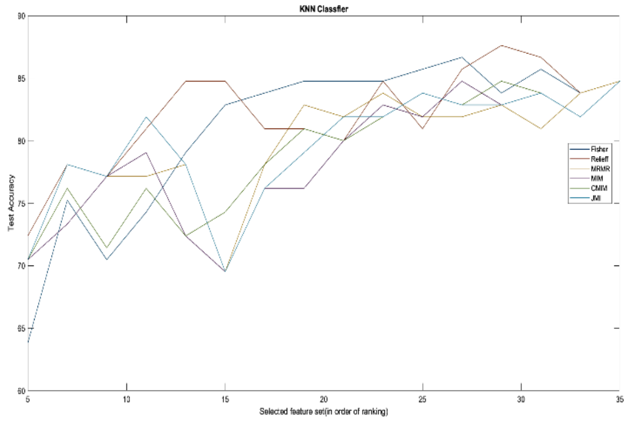

Fig.3. Training Accuracy Using KNN

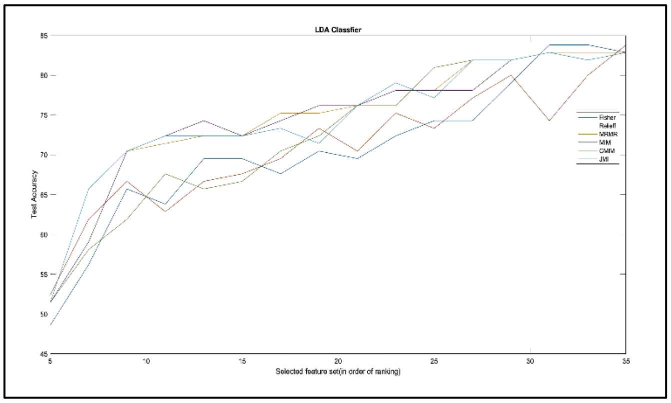

Fig. 4. Training Accuracy Using LDA

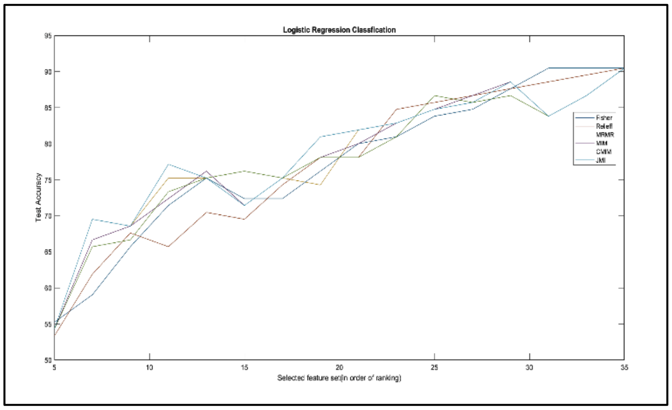

Fig.5. Training Accuracy Using LR

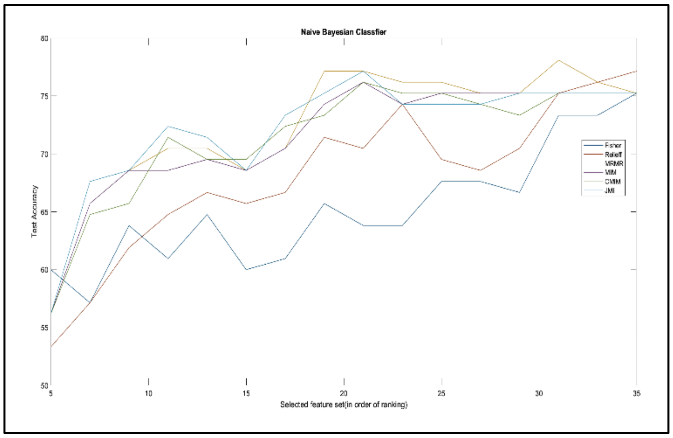

Fig. 6. Training Accuracy Using Naïve Bayesian

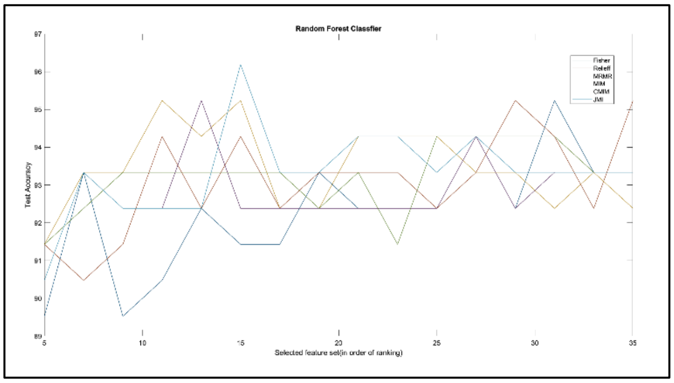

Fig.7. Training Accuracy Using Random Forest

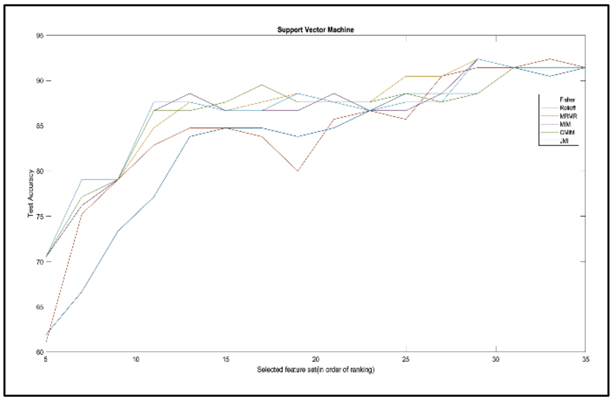

Fig. 8. Training Accuracy Using SVM

Conclusion

Emotion has a major part in our existence. Speech is a method of communicating with others. We have built a model in this paper to categories German speech signals into different emotional categories. This aims to improve the performance of the raw model of machine learning algorithms. The effectiveness of the categorization of emotions has been enhanced effectively utilizing the technique of feature selection based on common information and eliminates irrelevant characteristics from the retrieved feature set. The experimental results show the promising result with an accuracy of 96 percent for class emotional categorization.

References

O.Alsac,and B. Scott, “Optimal load flow with steady state security”, IEEE Transaction PAS -1973, pp. 745-751. May 1973.

D. Meyer, “Support Vector Machines”. FH Technikum Wien, Austria, August 5, 2015.

J. Naik and G. Doddington, ``Evaluation of a high performance speaker verification system for access control,'' in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2005, pp. 2392 2395.

M. G. Gomar, ``System and method for speaker recognition on mobile devices,'' Google Patents 9 042 867, Mar. 26, 2015.

M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, ``Speaker identification security improvement by means of speech watermarking,'' Pattern Recog- nit., vol. 40, no. 11, pp. 3027 3034, Nov. 2007.

B. M. Arons, Interactively Skimming Recorded Speech. Cambridge, MA, USA: Massachusetts Institute of Technology, 1994.

C. Schmandt and B. Arons, ``A conversational telephone messaging system,'' IEEE Trans. Consum. Electron., vols. CE 30, no. 3, pp. 21 24, Aug. 1984.

A. Maurya, D. Kumar, and R. K. Agarwal, ``Speaker recognition for hindi speech signal usingMFCC-GMMapproach," Procedia Comput. Sci., vol. 125, pp. 880 887, Jan. 2018.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, ``Speaker veri cation using adapted Gaussian mixture Models," Digit. Signal Process., vol. 10, nos. 1 3, pp. 19 41, Jan. 2000.

D. A. Reynolds, ``Speaker identi cation and veri cation using Gaussian mixture speaker models," Speech Commun., vol. 17, nos. 1 2, pp. 91 108, Aug. 1995.

W.-C. Chen, C.-T. Hsieh, and C.-H. Hsu, ``Robust speaker identi cation system based on two-stage vector quantization," J. Sci. Eng., vol. 11, no. 4, pp. 357 366, 2008.

D. B. A. Mezghani, S. Z. Boujelbene, and N. Ellouze, ``Evaluation of SVM kernels and conventional machine learning algorithms for speaker identi cation," Int. J. Hybrid Inf. Technol., vol. 3, pp. 23 34, Jul. 2010.

Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics 2001; 34(1):28–36.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In: Proceedings of the Acoustics Speech and Signal Processing (ICASSP) 2017 IEEE International Conference, pp. 2227-2231.

Nogueiras, A., Moreno, A., Bonafonte, A., & Marino, J. B. (2001). Speech Emotion Recognition Using Hidden Markov Models. In: Eurospeech 2001. Seehapoch, T., & Wongthanavasu, S. (2013). Proceedings of the 5th International Conference on Knowledge and Smart Technology (KST).https ://doi.org/10.1109/KST.2013.65127 93.

Terken, J. M. B. (1994). Fundamental frequency and perceived prominence of accented syllables. The Journal of the Acoustical Society of America, 95(6), 3662–3665.

Zhao, J., Mao, X., & Chena, L. (2019). Speech emotion recognition using deep 1D and 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312–323.

Zhao, J., Ma, R. L., & Zhang, X. (2017). Speech emotion recognition based on decision tree and improved SVM mixed model. Transaction of Beijing Institute of Technology,. https ://doi.org/10.15918 /j.tbit1 001-0645.2017.04.011.

Jean Shilpa, V., & Jawahar, P. K. (2019). Advanced optimization by profiling of acoustics software applications for interoperability in HCF systems. Journal of Green Engineering, 9(3), 462–474.

Jing, S., Mao, X., & Chen, L. (2018). Prominence features: Effective emotional features for speech emotion recognition. Digital Signal Processing., 72, 216–231.

Kakouros, S., & Rasanen, O. (2015). Automatic detection of sentence prominence in speech using predictability of word-level acoustic features. In: Proceedings of Inter speech, pp. 568–572.

Kakouros, S., & Rasanen, O. (2016). 3PRO an unsupervised method for the automatic detection of sentence prominence in speech. Speech Communication, 82(1), 67–84.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In Social media and machine learning. Intech Open. https ://doi.org/10.5772/intec hopen .84856 .

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. The Journal of the Acoustical Society of America, 118(2), 1038–1054.

Lieberman, P. (1959). Some acoustic correlates of word stress in American English. The Journal of the Acoustical Society of America., 32(4), 451–454.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE,. https ://doi.org/10.1371/journ al.pone.01963 91.

Mohammed Senoussaoui, Patrick J. Kenny, Najim Dehak, Pierre Dumouchel.(2011) “An i-vector Extractor for speaker Recognition with Microphone and Telephone Speech. ” CSAIL-MIT

Wei Li , Tianfan Fu and Jie Zhu.(2015) “An improved i-vector extraction algorithm for speaker verification Transient Flow.” EURASIP journal on Audio, Speech and Music Processing, pp 1-9

Najim Dehak, Patrick J. Kenny, Dehak, Pierre Dumouchel and Pierre Ouellet. (2011) “Front-End Factor Analysis for Speaker Verification.” Transactions on Audio, Speech and language Processing,vol.24,No.3

Padmanabhan Rajana b, Anton Afanasyeva, Ville Hautamakia and Tomi Kinnunena.(2014) “From single to multiple enrollment i-vectors: practical PLDA scoring variants for speaker verification. ” Digital Signal Processing

Dominic Mathew, V.D.Devessia and Tessamma Thomas. (2006) “A K-means Clustering Algorithm for Frequency Estimation and Classification of Speech Signals.” IEEE conference/ICSIP.

***