International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-3, May 2022
https://doi.org/10.55524/ijircst.2022.10.3.65
Article ID IRPP1214, Pages 406- 411
www.ijircst.org

A Review on Speech Emotion Recognition Using Machine Learning

Sk. Mohammed Jubear[1], D. Pavan Kumar Reddy[2], G. Subramanyam[3], Sk. Farooq[4], T Sreenivasulu[5], and N. SrinivasaRao[6]

[1,2,3,4,5,6] Department of Computer Science and Engineering, PACE Institute of Technology & Sciences, Vallur, Ongole, Andhra Pradesh, India
Correspondence should be addressed to Sk. Mohammed Jubear; 18kq1a05g0@pace.ac.in

ABSTRACT- This paper focuses on the development of a robust speech emotion recognition system using a combination of different speech features with feature optimization techniques and speech de-noising technique to acquire improved emotion classification accuracy, decreasing the system complexity and obtain noise robustness. Additionally, we create original methods for SER to merge features. We employ feature optimization methods that are based on the feature transformation and feature selection machine learning techniques in order to build SER. The following is a list of the upcoming events. A neural network can use either of these two techniques. As more feelings are taken into account, the feature fusion-acquired SER accuracy falls short of expectations, and the plague of dimensionality starts to spread due to the addition of speech features, which makes the SER system work harder to complete its task. This is due to the SER system becoming more complicated when voice elements are added. Therefore, it is crucial to create a SER system that is more trustworthy, has the most practical features, and uses the least amount of computing power possible. By using strategies that maximize current features, it is possible to streamline the feature selection process by reducing the total number of accessible choices to a more reasonable level. This piece employs a method known as Semi-Non Negative Matrix Factorization to lessen the amount of processing trash that the SER system generates. (Semi-NMF). This approach can be used to change traits that are capable of learning on their own.

KEYWORDS- Speech Emotion Recognition, Machine Learning, HCI, SER, MFCC

I. INRODUCTION

Speech emotion recognition (SER) is defined as the method of determining a person's mental state from their vocalizations. Everyday human interaction is heavily reliant on the expression of feelings. In the process of people sharing their perspectives with one another, their emotional states become apparent. Human-computer interface (HCI) experts are placing more and more emphasis on emotion recognition [1]. A person's ability to reason, focus, and solve problems is all affected by their mental condition. To effectively enhance human-computer contact, it is essential that systems be able to discern human emotions, and this is where affective computing comes in. (HCI). The experts found this to be a difficult topic as they tried to narrow their attention to it. Expression recognition can make use of a wide variety of signs, including facial data, bodily signals, vocal signals, and more. Several fields of study and practice make extensive use of mood recognition.

Many various kinds of information can be used to pinpoint emotional states. In terms of the numerous methods and sources that can be used to identify an individual's mental condition, speech signals are preferable to bodily signals like electrocardiograms. Reason being, getting the voice signal is easy and cheap. Communication through speech is the most natural and effective form of interaction between humans. Therefore, the idea of using the vocal signal as a means of communication in human-computer contact has evolved into the most effective and fastest way conceivable. Nevertheless, it's crucial for computers to be able to understand human communication. Over the span of the last few decades, there has been remarkable progress in the capacity of computers to understand human speech. This process is known as speech recognition. Speech, In order to identify a person or object, it is necessary to translate the spoken indicator into a series of phrases. Despite advances in voice detection, we have a long way to go before computers and humans sound as natural as each other when conversing. This is because a machine can't understand the speaker's emotional tone [2, 3]. To achieve this objective, it is essential to identify the emotions from the speaker's cues. Modern speech emotion recognition (SER) study has grown substantially for this reason. Using the information conveyed in a conversation, SER attempts to identify the speaker's mood. Depending on the context, human speech can communicate a wide range of emotions. It takes practice to pick up on the inflections of another person's speech and interpret them accurately. It can be difficult to figure out which traits work best for differentiating between various emotional states. The sounds will be different because individuals talk at various speeds, use a broader variety of phrase patterns, and interact in different ways. These factors have an instant effect on the speech traits used in SER, such as tone and energy [4]. Extracting characteristics from a voice signal that depict the

speaker's mental state is one of the most essential factors that must be given attention when delivering a talk.

Applications of Speech Emotion Recognition a variety of day-to-day applications make use of emotion identification in a variety of contexts. When it comes to human-computer interaction, feelings play a significant part. The goal of the emotion identification system is to intelligently categorize the range of transient feelings experienced by people after obtaining data on their utterances [1, 9]. The SER method is versatile and can be applied in a variety of contexts. In the field of medicine, it is the responsibility of the psychotherapist to determine the patient's mental health based on the counseling meetings. The psychological conditions of the patient include whether or not they have suicidal inclinations, whether or not they are depressed, and whether or not they have aberrant behavior. A verbal output can be used as the source of incoming data in the development of an expression identification system for these kinds of applications [10]. This can be accomplished by teaching a system with the speech data acquired from the psychotherapy meetings and identifying the human emotional state in greater detail. The SER system is a diagnostic tool that can be utilized by speech therapists, who specialize in the treatment of conditions related to the voice, speaking, and language. icSpeech is an example of this type of software, and it is designed to capture and evaluate speech patterns [11]. This is to determine whether or not the patient is under any sort of mental strain. In this investigation, the SER system identifies affective states based on prosodic, vocal tract, and glottal characteristics, which are considered speech features.

Customers are given the opportunity to provide comments and ask questions about specific products through the services provided by contact centers. To increase their number of sales, the businesses that produce these goods need to pay careful attention to the level of gratification they provide to their clients while they are providing these services. However, it frequently has difficulties in resolving the disputes that its consumers have. As a result, the representatives who work in customer service need to be educated to resolve grievances while exhibiting a great deal of tolerance. In light of these considerations, a methodology has been developed to evaluate the level of consumer satisfaction by taking into account captured phone conversations. On the other hand, this procedure takes place after the contact has been attended to. An SER system that uses intonation, energy, and rate of speech characteristics to identify feelings such as frustration can be developed for use in real-time evaluation with the purpose of analyzing the behavioral state of consumers such as the level of exasperation they are experiencing. This results in a higher standard of service provided by the contact assistant. In addition, the system that was developed can also be used to determine the customers' states of mind, such as fear or wrath, in order to determine the level of severity of the situation and prioritize the customers' calls. Because of this characteristic, those who work in emergency services will have a much better chance of avoiding accidents. Lie detectors can be very useful tools in the course of criminal investigations, particularly when it comes to determining whether or not individuals are being dishonest. A lie detector can assist in determining whether or not a person is telling the truth when they are conversing. The lie detector is utilized in the Central Bureau of Investigation (CBI) for the purpose of locating offenders and also for the purpose of preventing misconduct in the Cricket Council [12]. X13-VSA PRO COBRA Voice The Lie Detector is an innovative, cutting-edge, and technologically advanced computer software device. It is a tension detector that can quickly determine the truth from the sound of a person's speech [9]. If an automated teller machine in the financial industry is designed with the one-of-a-kind capability of integrating speech recognition, speaker identification, and expression recognition, then it will be capable of maintaining a high degree of security while simultaneously accessing confidential information. During the registration process for new customers, the system may record their sounds for the purposes of verifying their identity and determining the extent to which their speech exhibits indications of wrath, anxiousness, or dishonesty [13]. Therefore, if there is any kind of deception, the automated teller machine will not distribute the currency but will instead restrict the ATM card in order to provide security.

II. SPEECH PRE-PROCESSING

To maximize the effectiveness of the feature extraction module, the voice signal is pre-processed before being sent to it. Filtering, framing, and windowing are preprocessing steps. After the voice signal has been preprocessed [23], the physical variables such as tone, energy, and formants can be extracted. When a voice stream is filtered, the noise introduced by external disruptions or by the recording process itself is diminished. To compensate for the attenuation of higher harmonics in the speaking stream as it is produced by the vocal tract, a pre-emphasis filter is used. Non-stationary signs, like the voice signal, are notoriously challenging to study. Because of this, it is helpful to divide the voice stream into the same number of samples as the number of segments. The technique of feature extraction informs the decision regarding frame size. Because of this, a gap between the frames is permitted. There are breaks in the incoming data signal at the frame boundaries because of how the signal is split into frames. A curved window is applied to each frame to remove the gap that would otherwise appear. There are numerous types of windows, such as Hamming, Hanning, Rectangular, Barlett, Kaiser, etc.

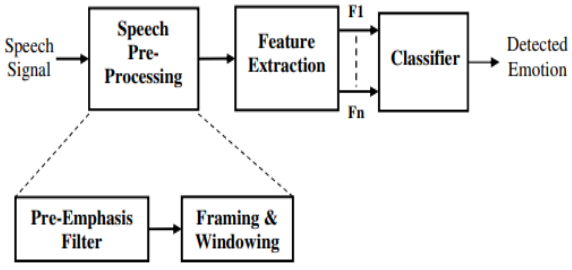


Figure 1: Basic Speech Emotion Recognition System
Feature Extraction

The process of feature extraction seeks to reduce the dimensionality of the features used to analyze speech in order to derive information about the speaker's feelings. (F1 - Fn). Expression recognition relies heavily on the linguistic features that can be used to isolate the speaker. Methods for collecting speech characteristics are used in the area of speech emotion recognition to better determine the speaker's emotions from the data they produce. Numerous speech characteristics have been studied in an effort to identify speaker emotion, but a definitive set of features has yet to be identified. [8] Speech features that are affected by an individual's emotional state include qualitative, spectrum, continuous, and Teager Energy Operator (TEO) based features. A speech expression's emotional content has the greatest impact on its continuous prosodic features, such as tone, zero-crossing rate, and vitality. Energy, articulation rate, spectral information, and basic frequency are all aspects of speaking that fit into this category. (f0). The way someone speaks has a significant impact on how they are made to feel. Structures in the form of feature boundaries, vocal levels, voice tone, and time intervals can be identified. Features obtained from spectral analysis are presented as a real-time snapshot of the conversational flow. The distribution of a statement's spectral energy depends on its emotional content. It has been noticed that the energies of words spoken while experiencing high-arousal emotions, like joy or anger, are concentrated at higher frequencies, while the energies of words spoken while experiencing low-arousal emotions, like sadness, are concentrated at lower frequencies. The most common spectral-based features used in mood detection are the Mel Frequency Cepstral Coefficients (MFCC) and the Linear Predictive Coefficients (LPC). Speech is produced by the non-linear breathing process of the vocal tract. Muscular tension in the mouth affects the flow of air through the vocal tract's sound-producing system, and this impact is intensified under pressure. Non-linear speech traits are thus crucial to the process of speech recognition. Considering that hearing is already the most common means of energy tracking, Teager and Kaiser reasoned that it should serve as the foundation for the Teager Energy Operator role. The use of multiple speech traits simultaneously has replaced previous methods as the gold standard for speech emotion recognition.

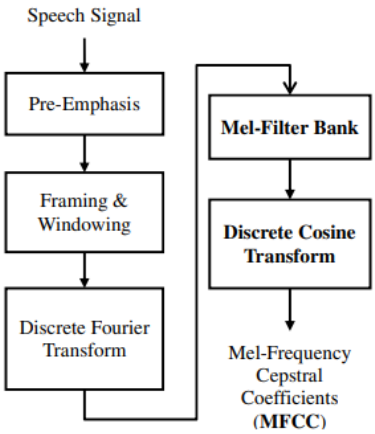


Figure 1: MFCC Feature Extraction

Speech Emotion Recognition using Spectral and Teager Energy Feature Fusion Majorly, the detection of stressed emotions in speech plays a vital role in the real-time applications like the mood of a car driver to avoid accidents, a student's mental state to give them proper counselling, a child's psychological state to improve their parents and other acquaintance, etc. and also in aircraft cockpits the speech recognition systems trained with stressed speech provided better results compared to normal speech [4]. Helping anxious people early can prevent many problems. Researchers previously tried to identify all emotions, but they ignored the most evident ones like rage, fear, sorrow, disdain, annoyance, etc. The TEO feature, which detects emotions, can be used for forced voice emotion detection. Using TEO traits with spectrum ones improves mood detection.

III. TEAGER ENERGY OPERATOR (TEO)

When the speech signal is generated, the speech produced under stressful circumstances has an effect on the irregular movement of air in the system that comprises the vocal tract. As a result, the recognition of speech relies heavily on the utilization of these non-linear aspects of speech. On the basis of his experiments, Teager suggested an energy operator, also known as the Teager Energy operator [8]. This energy operator is a measure of the energy contained in speaking signals. Teager demonstrated, through his experiments, that the passage of air through the vocal tract is segmented, and that it subsequently follows the sides of the vocal tract. Teager came up with the structure of the vocal tract and modeled the process of speech generation based on his observations of the outcomes of a few whistle experiments. These findings are depicted in Figure 3.2. According to this concept, air leaves the glottis in the form of a stream and connects itself to the wall of the vocal tract that is closest to it. The cyclones of air are created when air is transferred between the true vocal folds and the false vocal folds through the chamber of the vocal tract. The majority of the air that is transmitted through the vocal tract does so near the mouth as it travels along the sides of the vocal tract. The movement of the cyclone is the component of this model that should be given the most attention. Only the glottis, which is the only part of the vocal tract that has been constructed, is involved in the generation of sound in the conventional paradigm of how speech is made. Teager, on the other hand, claimed that vortices in the region of the artificial vocal folds also actively generate sound, which produces modulations in the speaking signal.

Proposed SER System using Semi-Non Negative Matrix Factorization (Semi-NMF) Speech preprocessing, feature extraction, and categorization are the only three phases that make up a conventional SER system [8]. The majority of the currently available SER systems perform emotion identification with the help of a combined collection of speech characteristics. This adds more work to the categorization model's already extensive computational burden. As can be seen in Figure 2, the semi-NMF optimization method is incorporated into the process of developing the SER system before the classification of the characteristics that are used to acquire the feelings in order to surmount this

shortcoming. In the system that has been suggested, following the extraction of features but prior to the categorization of feelings, a semi-NMF with SVD initialization feature optimization technique is utilized in order to cut down on the original extremely large feature sets.

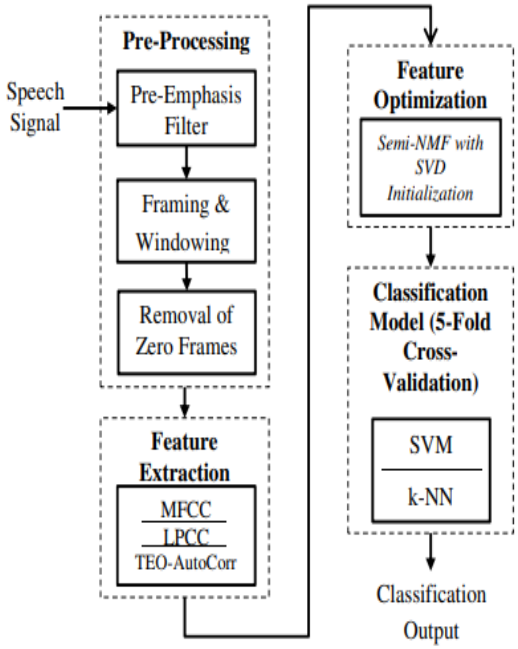


Figure 3: Proposed Speech Emotion Recognition System using Semi-NMF

IV. NOISE ANALYSIS

The accuracy of the SER suffers when there is a lot of surrounding noise because the voice stream is skewed. Therefore, it is necessary to create a SER system that can function in noisy environments. Improved voice emotion detection ability is the focus of this thesis, which makes use of noise-resistant Power Normalized Cepstral Coefficients (PNCCs). (SER). This works to counteract the sometimes obtrusive effects of noise during emotional speech detection. Moreover, NMF-based voice de-noising is applied in advance of SER to enhance the accuracy of the latter. This is done to achieve noise resistance in low SNR environments. When compared to both the standard (which does not account for feature optimization or feature selection) and the works that have been previously published in the pertinent literature, all of the freshly created SER systems perform admirably. The SER system was developed using a number of different categorization strategies, including the Gaussian mixture model, k-Nearest Neighbors, and Support Vector Machines, as well as the hold-out and cross validation approaches. Aiming to classify feelings, these methods were employed.

When working in real time, the speech stream is vulnerable to different types of background noise. This is a major roadblock to the SER system's effectiveness. One way around this shortcoming is to try mood detection after background noise has been removed from the voice. To date, numerous speech de-noising methods have been created to enhance speech intelligibility without introducing unwanted disturbances [9], [8]. The purpose of speech de-noising methods is to get rid of any and all background noise that may have contributed to the distortion of the original speech. One of the easiest and most frequently used methods is spectral subtraction, which is used to determine the noise spectrum signature in the amplitude domain. The technique entails removing the sound of speech from the background din. In contrast, the spectral reduction method can only be used to analyze motionless or nearly steady data. In the setting of blind source separation [10], Non-negative Matrix Factorization (NMF), a method that uses voice signal models, has lately become the predominant technique that is used. The spectral amplitude (denoted by H) is factored into two non-negative matrices (H = UV) using this method. The columns of the 'U' matrix indicate the amplitude spectral patterns, and the rows of the 'V' matrix reflect the gain factors.

V. SIMULATION RESULTS AND PERFORMANCE EVALUATION

The proposed speech recognition (SER) system sifts through the audio data in search of 1602 INTERSPEECH Paralinguistic and GTCC features. This vast collection of characteristics is fed into the UFSOL and FSASL algorithms to help determine which ones are most important. In this research, we use the support vector machine (SVM) classification technique to label user mood using Linear and Radial Basis Function (RBF) kernels, Hold-Out, and 10-fold Cross-Validation. At first, the collection of spoken-word signals is split into a training dataset and a testing dataset. With hold-out validation, only a subset of the data is put to use for evaluation, while the leftover 80% is put to good use in the learning process. The proposed SER system was taught with the 10-fold cross-validation method, and its accuracy was evaluated afterward. Therefore, the full dataset is randomly split into ten pieces, with nine pieces used for training the classifier (SVM) and the tenth piece serving as the hold-out data or test data, which is used for testing. The training process is repeated 10 times, or "folds," to ensure that all data in the collection is properly trained. The term "folding" describes this procedure. In order to conduct the experiments, we make use of the EMO-DB and IEMOCAP tools. The classification accuracy is used as a machine learning performance metric to assess the effectiveness of the proposed SER system. This is important for evaluating the system's efficiency. Ten-fold cross-validation is used to hone the accuracy of the proposed SER system, and Hold-Out Validation is used to ensure it holds up in practice. These two approaches were employed. An Intel(R) Xeon(R) CPU E3-1220 v3 operating at 3.10 GHz with 64-bit support and 16 GB of random access memory is used to perform all of the computations. (RAM).

Table 1: Simulation Parameters of Proposed SER System using unsupervised FS

Parameters	Specifications
Pre-Emphasis Filter	Coefficient, a = 0.97
Frame Size/ Length	4096 samples
Frame Overlap	1024 samples
Type of Window	Hamming
Gammatone Filter	Filter order = 4 $f_{low} = 62.5Hz$, $f_{high} = 3400Hz$ Number of gammatone filters = 20
Validation	Hold-Out Validation (80/20) 10-Fold Cross-Validation
Support Vector Machine	Kernel = Linear, Radial Basis Function

To select the first prominent features which give the highest accuracy, to select the initial feature set, the feature selection matrix of both UFSOL and FSASL algorithms are given to the SVM classifier as shown in Figure 3.

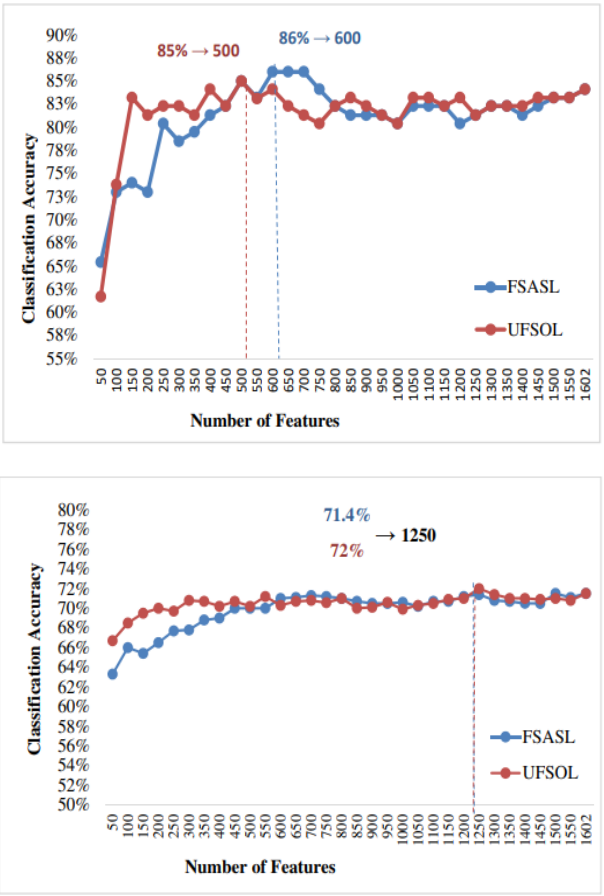
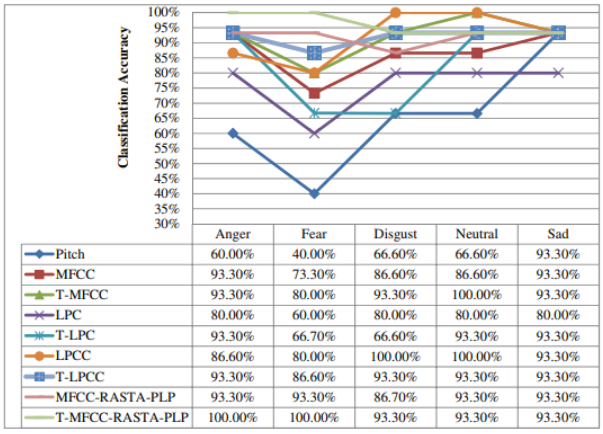


Figure 4: Variation of classification accuracy in proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB.



	Anger	Fear	Disgust	Neutral	Sad
Pitch	60.00%	40.00%	66.60%	66.60%	93.30%
MFCC	93.30%	73.30%	86.60%	86.60%	93.30%
T-MFCC	93.30%	80.00%	93.30%	100.00%	93.30%
LPC	80.00%	60.00%	80.00%	80.00%	80.00%
T-LPC	93.30%	66.70%	66.60%	93.30%	93.30%
LPCC	86.60%	80.00%	100.00%	100.00%	93.30%
T-LPCC	93.30%	86.60%	93.30%	93.30%	93.30%
MFCC-RASTA-PLP	93.30%	93.30%	86.70%	93.30%	93.30%
T-MFCC-RASTA-PLP	100.00%	100.00%	93.30%	93.30%	93.30%

Figure 5: Most reliably rage

Figure 5 shows that rage is the most reliably identified feeling (93.3%), followed by fear (20%), contempt (80%), indifferent (46.6%), and melancholy (60%) when using the pitch feature extraction method for male speech. 100% precision is achieved for rage, 86.7% for dread, 60% for contempt, 93.3% for indifferent, and 80% for sadness using MFCC. Anger and sadness are correctly recognized at 100%, dread at 86.6%, revulsion at 80%, and indifferent at 80% using LPCC. For the feelings of rage and sadness, as well as dread (93.3%), contempt (86.67%), and apathy (80%), the MFCC-RASTA-PLP method achieves perfect precision. As a result of integrating these enhancements into TEO, the system becomes more precise than its predecessors.

VI. CONCLUSION

In this study, we combine the Paralinguistic and GTCC features that were presented at INTERSPEECH 2010 by using the unconstrained feature selection techniques UFSOL and FSASL to identify the optimal combination. On top of the UFSOL and FSASL techniques, a novel SuFS algorithm has been introduced. The goal of this algorithm is to further reduce the feature dimension while acquiring the same level of performance in the recommended SER system. An SVM classifier equipped with linear and RBF kernels is utilized in conjunction with the EMO-DB and IEMOCAP datasets in order to evaluate the efficiency of the proposed SER system. The problem of over-fitting is avoided by training the feature sets using a 10-fold Cross-validation scheme. Additionally, the effectiveness of the recommended SER system is evaluated with new data using a Hold-Out validation scheme. The recommended SER system for EMO-DB data provides the highest classification accuracy (86%) with SVM equipped with a Linear kernel, followed by FSASL (86%), UFSOL (85%), and FSASL-SuFS (85%). Likewise, the SVM classifier with the RBF kernel achieves the highest classification accuracy for the IEMOCAP database, which is 77%. This is followed by FSASL-SuFS, which achieves 69%, and the remaining approaches, which achieve 69%. According to the findings, the proposed SER system is not only superior to the benchmark, which is another SER system that does not include feature selection, but it is also superior to the state-of-the-art system. (literary works that have already been published).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," Neural Comput. Appl., 2000.

[2] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," J. Pers. Soc. Psychol., 1996.

[3] M. J. Kim, J. Yoo, Y. Kim, and H. Kim, "Speech emotion classification using treestructured sparse logistic regression," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015.

[4] S. Lukose and S. S. Upadhya, "Music player based on emotion recognition of voice signals," in 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017, 2018

[5] S. Ramakrishnan, "Recognition of Emotion from Speech: A Review," in Speech Enhancement, Modeling and Recognition- Algorithms and Applications, 2012.

[6] D. A. Cairns and H. L. John Hansen, "Nonlinear analysis and classification of speech under stressed conditions," J. Acoust. Soc. Am., 1994.

[7] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., 2011

[8] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," Biomed. Signal Process. Control, 2015.

[9] PrasaduPeddi (2019), Data Pull out and facts unearthing in biological Databases, International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32

[10] Z. W. Huang, W. T. Xue, and Q. R. Mao, "Speech emotion recognition with unsupervised feature learning," Front. Inf. Technol. Electron. Eng., 2015.

[11] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," Neural Comput. Appl., 2014.

[12] PrasaduPeddi (2018), "A Study For Big Data Using Disseminated Fuzzy Decision Trees", ISSN: 2366- 1313, Vol 3, issue 2, pp:46-57.

[13] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., 2008.

[14] X. Valero and F. Alias, "Gammatonecepstral coefficients: Biologically inspired features for non-speech audio classification," IEEE Trans. Multimed., 2012.