Journal of Positive Sciences (JPS) **eISSN: 2582-9351, https://positive-sciences.com** **JPS. 2023 Sep; 3(5): 39-48**

Article info **Research Article**
Received on: 24.08.2023
Accepted on: 24.09.2023
Published on: 30.09.2023
doi: https://doi.org/10.52688/ASP56048

Classification of speech recognition by using sequential minimal optimization algorithm

Ali Najdet Nasret [1,*]

[1] Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq

[*] alinajdet@ntu.edu.iq

ABSTRACT

The categorization and recognition is a very recent development in the realm of machine learning. This study shows the categorization of emotions using the architectural framework of a Distributed Speech Recognition System (DSRs), accompanied with the related results of performance evaluation. The temporal patterns of semantic units, such as sentences and words, characterized by using a set of 3800 statistical factors. The use of the KDDM (Knowledge Discovery and Data Mining) program was employed to conduct the procedure of determining the most pertinent components for classifying emotional states. Subsequently, a thorough analysis was performed on the data obtained from various classification methodologies. The findings, derived from the analysis of the California Database of Emotional Speech and the Actual Stress corpus and Speech Under Simulated, indicate that the optimal outcomes are attained by employing a Sequential Minimal Optimization (SMO) algorithm to feeding and training the Support Vector Machine (SVM). The aforementioned result is achieved by the normalization and discretization of the statistical parameters given as input.

Keywords: Sequential Minimal Optimization, KDDM, Support Vector Machine, Speech Recognition

INTRODUCTION

The presence of covert affective information inside speech is of considerable importance in human communication and interaction, since it provides feedback without altering the linguistic content. In the context of interpersonal communication, it is beneficial to discern and distinguish between two separate channels that exist within spoken conversation. The primary conduit, which is linked to the syntactic-semantic module, is tasked with conveying language information. In contrast, the secondary channel fulfills the function of conveying paralinguistic signals, including elements such as intonation, emotional state, and non-verbal movements [1][2]. Upon being acknowledged and processed, the supplementary information provided via the secondary channel serves to improve the efficacy of various tasks. These functions encompass the convergence of speech patterns among interlocutors, the ability to interact with the primary channel to discern nuances such as humor or rhetorical questioning, the enhancement of evaluative discernment to detect falsehoods, the prevention of misinterpretation by emphasizing key elements of a sentence, and the provision of supplementary details regarding the speaker's attributes such as geographical origin, gender, or age. A multitude of designations are often used in colloquial discourse to denote the diverse range of emotional states [3][4]. Plutchik and Whissel have identified a total of more than 120. At now, it is challenging to conceive of an artificial system that has the capability to achieve such a significant level of differentiation. In order to surmount this challenge, it is possible to depict emotional states within a continuous space that encompasses two or three dimensions. One methodology entails the allocation of valence to individual coordinates, which serves to denote the positive or negative characteristic of the emotion. Furthermore, it is possible to quantify the degree of activation in order to measure the amount of enthusiasm shown by the speaker. Finally, the degree of dominance may be assessed in order to evaluate the speaker's level of submissiveness or assertiveness. The capacity of a system to discern emotional states from speech has significant promise for diverse applications across severalareas [5][6]. Domains include a wide variety of fields, from psychiatric diagnosis to the toy industry to home jukeboxes to CRM to automatic Speech Recognition to automated learning to Speech Synthesis to voicemail systems and alarms. Some possible definitions of a spoken emotion recognition system include:

***Corresponding author**
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

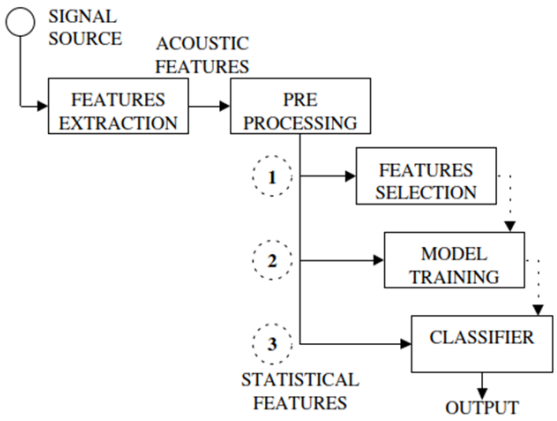


Figure 1: The provided visual representation depicts the system of categorization

An automatic categorizing system of any type. Figure 1 presents a comprehensive block diagram showcasing the three basics components crucially engaged in the process of information gathering, specifically focusing on the audio signal. The aforementioned components include the extraction and processing of parameters, together with the categorization of semantic units. The procedure has three discrete stages [7][8]. The first two stages of the process are devoted to the establishment of the system, whereby statistical characteristics are found during the first phase, and models are then trained for the purpose of classification during the second phase. On the other hand, step 3 employs the detected statistical traits and trained models to categorize semantic units with unfamiliar emotional contents. The present study used a phase 3 technique to investigate utterances containing identifiable emotional content, aiming to evaluate the effectiveness of the system in accurately classifying such phrases [9][10][11]. The audio data were used to generate a complete collection of 4000 statistical characteristics for every semantic unit. In order to enhance the process of constructing a classification system, it is important to possess speech samples that include a wide range of emotional states, including those that may be categorized as one of the key emotions under investigation. The speech corpora were used for both the training and assessment stages.

EXTRACTION OF FEATURE

The feature extraction procedure conforms to the guidelines outlined by the field of voice recognition. The standard pertains to the calculation of feature vectors derived from speech waveforms that have been sampled at a frequency of 17 kHz. The input signal, which does not include any offset, is segmented into frames of N = 550 samples. Each frame overlaps with both the preceding and succeeding frames. The duration of each frame is 25 milliseconds. The frame shift interval, denoted as M, is equal to 160 samples, which corresponds to a duration of 10 milliseconds. The recovered feature vector for each frame has 15 coefficients, including the log-energy coefficient, the 13 cepstral coefficients C1 to C13, the voicing class and pitch period. The logarithmic energy coefficient and the 13 Mel-frequency cepstral coefficients (MFCCs) are subjected to calculations for their first and second derivatives. The pitch period is further used for the computation of jitter, a metric that quantifies the variation in basic frequency from one period to another. The calculation of jitter is performed by applying a formula to the successive voiced periods.

$$Jitter = \frac{\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i}$$

The equation presented above defines the variable *T*" as the pitch period of the ith window, while N denotes the total number of voiced frames inside the segment. In all, there are 41 time series per segment, excluding the classification of the frame. The result is obtained by doing the multiplication of 13 by 3 and afterwards adding 2. Prior to further processing, any null elements (referring to frames without speech for which the front-end is unable to determine the pitch) are eliminated from the pitch sequence. Furthermore, the MFCC sequence is subjected to the removal of the start and final frames that are identified as possessing quiet. This process successfully eliminates any frames of stillness at the beginning and end of the segment. The attribution of an individual's emotional state only based on a single measure extracted from a brief temporal window lasting few milliseconds has no academic relevance. Conversely, it is crucial to analyze the temporal trajectory of the parameter. The temporal components of vocal signals are subjected to analysis in order to derive certain statistical parameters, such as the mean, minimum, and maximum values. The estimated sequences are derived from the observed patterns shown by each of the retrieved characteristics. The sequences encompass various elements. Statistical data is calculated for each of these sequences. The terms to be examined within the scope of this academic discourse encompass the following: mean, variance, maximum value, minimum value, range (defined as the difference between the maximum and minimum values), first quartile, second quartile (commonly referred to as the median), third quartile, and interquartile range (computed as the disparity between the third quartile and the first quartile). Two more statistical features are obtained by calculating the ratio of the count of relative minima to the count of

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

frames, as well as the ratio of the count of relative maxima to the count of frames. The categorizing of frames yields additional sequences, including the length of silent segments, the duration of unvoiced segments, the duration of mixed segments, and the duration of voiced sections. All the statistical information described above has been calculated for these recently created sequences. An additional statistical feature that is calculated involves the calculation of the ratio of transitions between distinct states to the total number of frames in the segment. Consequently, the total number of characteristics for each segment to be classified was calculated as $41 \cdot 10 \cdot 9 + 41 \cdot 2 + 4 \cdot 9 + 1$, resulting in a value of 3809.

WORD DATABASES

The compilation of emotional speech data sets is undeniably valuable for academics with an interest in the identification of emotional speech. The current state of research on emotional speech recognition reveals a restricted focus on specific emotions [12]. This limitation arises from the fact that the bulk of emotional speech data sets typically include just 5 or 6 emotions, despite the existence of several other emotion categories in real-world contexts. The present research used two voice corpora as the primary source of data acquisition. The California Database of Emotional Speech is the first corpus, and it is composed of semantic units such as phrases and written in Spanish. The second corpus, referred to as Speech Under Simulated and Actual Stress, is written in the English language and consists of semantic units that are composed of individual words.

MOODS IN SPEECH: A CALIFORNIA DATABASE

The database consists of six fundamental emotions, including anger, weariness, contempt, anxiety, happiness, and sorrow, in addition to neutral speech. A group of ten proficient Spanish actors, consisting of an equal distribution of five female and five male individuals, were engaged in the task of simulating various emotions. As a result, they generated a total of ten utterances, including five brief and five extended lines. These sentences were carefully crafted to be applicable in ordinary conversations and were designed to be comprehensible across a range of emotional contexts. The evaluative process included the analysis of about 800 phrases, including seven distinct emotions, ten different actors, and ten sentences per actor, along with additional alternate versions. The primary focus of the evaluation was to assess the recognizability and naturalness of the recorded speech material. This was achieved by a forced-choice automated listening test, which involved the participation of a panel of 20-30 judges. Following the process of selection, the database included a cumulative count of 494 sentences, with 286 of them being articulated by women and 208 by males. The distribution of phrases across different emotional states was uneven, with 55 statements expressing fear, 38 sentences expressing contempt, 64 sentences expressing happiness, 79 words expressing weariness, 78 sentences expressing neutrality, 53 sentences expressing melancholy, and 127 sentences expressing anger. In addition to being categorized for the categorization of the seven unique emotional states (referred to as 7EMOTIONS), the sentences inside the database were also organized in a manner that allows for differentiation among the various groupings of states.

EmoNoEmo: (neutral)- (melancholy, happiness, fear, repugnance, weariness, anger)

Evaluation: (neutral)-(happiness)- (melancholy, anxiety, repugnance, weariness, anger)

Activation: (neutral)- (weariness, melancholy)- (happiness, fear, repugnance, anger)

THE IMPACT OF STRESS ON SPEECH PERFORMANCE IN SIMULATED AND REAL- LIFE SITUATIONS

The database has been divided into five distinct categories, which together contain a diverse range of pressures and emotions. The five stress domains encompass various aspects of human communication and behavior. These domains include talking styles, which encompass different speech characteristics such as speed, volume, tone, and emotional expression. Another domain involves the performance of a single tracking task or the production of speech in a noisy environment, which is known as the Lombard effect. Additionally, there is a domain that involves engaging in a dual tracking computer response task. Another stress domain involves subjecting individuals to tasks that induce actual physical or emotional stress, such as experiencing high G-forces, the Lombard effect, exposure to noise, or inducing fear. The realm of psychiatric analytic data primarily centers on the examination of speech patterns shown during periods characterized by melancholy, fear, and anxiety. The database consists of two categories of speech recordings: simulated speech under stress, referred to as the Simulated Domain, and actual speech under stress, referred to as the Actual Domain. The SUSAS database comprises a collection of 35 aviation communication phrases that are often misinterpreted, forming a comprehensive vocabulary set. The aforementioned utterances are articulated by a collective of nine male people. Every style includes a pair of recordings with the repetition of the same phrase by both speakers. The audio is captured at a sampling frequency of 8 kHz, with each sample being encoded using a precision of 16 bits. The present investigation included a range of sonic stimuli, including anger, speed, the Lombard effect, interrogative intonation, slow speech rate, and soft volume. The need of include a 110 ms interval of quiet at the beginning of each recording was determined due to the limited length of the recordings, hence hindering the accurate extraction of features by the front-end system.

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

THE PROCESS OF SELECTING, DISCRETIZING, AND TRANSFORMING FEATURES

THE PROCESS OF SELECTING FEATURES

While it may seem natural that a classification system's discriminating skills would be enhanced by a larger number of characteristics, empirical investigations have shown that this is not necessarily the case. The system experiences improvements in efficiency and speed by lowering the size of the classification vector, resulting in a more concise and comprehensible dataset for interpretation by the learning algorithm. The primary approaches for feature selection may be categorized into two groups: wrapper methods and filter methods. Wrapper approaches include the interaction between the classification algorithm and the set of components. These methods are known for their higher accuracy, albeit they do demand a longer processing time. Filter techniques are considered to be independent, since they do not need any contact with the classification algorithm. Consequently, they are able to operate at a quicker pace. Filter models are a favorable option in situations when there is a substantial volume of training data due to their computational efficiency and their ability to operate independently of the learning method. One approach used to reduce superfluous and inconsequential elements is the identification of components that exhibit strong correlation with a certain class, however have weak correlation between themselves. The analysis can be conducted using forward selection, which involves initially having an empty list and iteratively adding new attributes until the improvement in performance falls below a predetermined threshold. Alternatively, backward elimination can be employed, where the analysis begins with a vector containing all components and gradually eliminates the least favorable attributes. In addition, there are more intricate search techniques, such as the best-first approach. This method maintains a comprehensive record of assessed component subsets, arranged based on performance metrics, enabling the possibility of revisiting prior configurations. Natural selection is the underlying idea behind genetic algorithms, which provide a search mechanism based on this principle. In this experiment, the CFSSubsetEval feature selection approach, which is offered by the KDDM software, was used. The algorithm employs Correlation-based Feature Selection as a methodology for evaluating features, aiming to detect and eliminate components that have substantial relationships with each other. In order to identify the most appropriate subset, we used a best-first search approach in conjunction with a stratified 10-fold cross-validation technique. Consequently, a grand total of 10 unique sets of predetermined parameters were obtained. In the domain of classification, it is feasible to choose a parameter x for the purpose of classifying data inside each of the 10 subgroups. However, it should be noted that parameter y is selected for classification in only nine out ofthe ten subgroups. Furthermore, it is important to note that an additional parameter, denoted as z, may only be selected for classification inside a single subset out of the 10 available subsets. This trend persists for more factors as well. A collective set of parameters designated as "1-10" has been built based on the selected parameters for the various subgroups. This set includes all the parameters that have been chosen for classification in at least one of the subgroups. Furthermore, we have developed the "2-10" composite, which encompasses characteristics that have been chosen for classification in a minimum of two of the subgroups. The aforementioned procedure continues until the "10" aggregate, which only consists of parameters that have been chosen for classification in all ten of the subsets that have been considered.

THE PROCESS OF DISCRETIZING FEATURES

In some instances, certain classification methods may not be well-suited for properly managing the continuous nature of the feature vector components presented. To enhance the efficiency of these algorithms, it would be advantageous to discretize the continuous domain of the different components. This research used a variety of methodologies to investigate the potential improvement in performance that may be achieved via the classification of emotional states.

NORMALIZATION FEATURES

Additional features may be produced from the original dataset to better describe the data for the learning technique used. Data normalization is a widely used method for accomplishing this goal. The main aim of frequently used data normalization procedures is to guarantee that all components are confined inside a pre-established range. The current study used both the z-score and max-min techniques. The aforementioned statement affects the original scope.

$$[A_{max}, A_{min}] \text{ Expanding the existing scope } [\hat{A}_{max}, \hat{A}_{min}]$$

$$\hat{v} = \frac{v - A_{min}}{A_{max} - A_{min}} \left(\hat{A}_{max} - \hat{A}_{min} \right) + \hat{A}_{min}$$

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

The second shifts

$$\hat{v} = \frac{v - A_{mean}}{A_{std} - A_{dev}}$$

THE FINDINGS OF THE STUDY ARE AS FOLLOWS

Among the several categorization algorithms offered by KDDM, our original preference was to use those that did not need excessively lengthy processing durations. The aforementioned classifiers include NaiveBayes, NaiveBayesSimple, J48, RandomForest, REPTree, Nnge, Part, Ridor, RBFNetwork, SimpleLogistic, SMO, IB1, Hyperpipes, and VFI. The NaiveBayes algorithm is a commonly used probabilistic classifier based on the Naive Bayes theorem, while NaiveBayesSimple is a simplified version of the Naive Bayes algorithm. The J48 algorithm is a decision tree learner that is derived from the C4.5 algorithm. The RandomForest algorithm is used to generate random forests for the purpose of categorization. The REPTree algorithm is a rapid tree learning method that incorporates the technique of reduced-error pruning. The NNGE algorithm is a method based on nearest-neighbor principles, which is used to produce rules via the utilization of nonvested generalized exemplars. The rules that are created from partial decision trees are acquired via the use of the J48 method. The Ridor algorithm employs the ripple down approach for the purpose of rule learning. The responsibility of implementing a radial basis function network lies with the RBFNetwork class. The SimpleLogistic technique is designed to build linear logistic regression models with the added functionality of attribute selection. The Sequential Minimum Optimization (SMO) algorithm is a widely used technique utilized in the context of support vector classification. It is based on the principles of minimum optimization and is designed to efficiently solve the support vector classification problem. The IB1 algorithm maybe classified as a fundamental instance-based learner that utilizes the nearest-neighbor approach. Hyper pipes is a very efficient and adept learner that depends on hypervolumes inside the instance space. The voting feature intervals (VFI) technique is a straightforward and efficient approach. All algorithms under investigation were executed with the default settings, except for the IBk algorithm, which had its parameter k manually set to a value of 2. To that end, we started by analyzing the EMO-DB speech corpus to establish which approach is most productive. When the classification systems were originally put into use, they made use of 3809 of the available characteristics. With an average recognition rate of around 79%, the SMO algorithm demonstrated exceptional performance given the current circumstances. Subsets of features were identified using the procedures outlined in Section 4.1, and their effectiveness in evaluating the performance of several classifiers was analyzed. Figure 2 shows how well various classifiers perform in terms of the typical percentage of recognition for the seven distinct emotional states that make up the database. This evaluation is based on the use of various aggregates of metrics. The experiment included doing tests on the same dataset that was used for training purposes, using a technique known as "stratified cross-validation," specifically employing a 10-fold cross-validation approach. The training process included using 90% of the database, while the remaining 10% was allocated for testing purposes. The aforementioned process was iterated a total of ten times, whereby each iteration had a distinct split of 90% and 10% respectively. Upon study of the acquired findings, it is evident that the different classification methods exhibit non-uniform performance as the number of factors used varies. The findings obtained from the 1-10 scale to the 5-10 scale are classified into three distinct groups. The first category include.

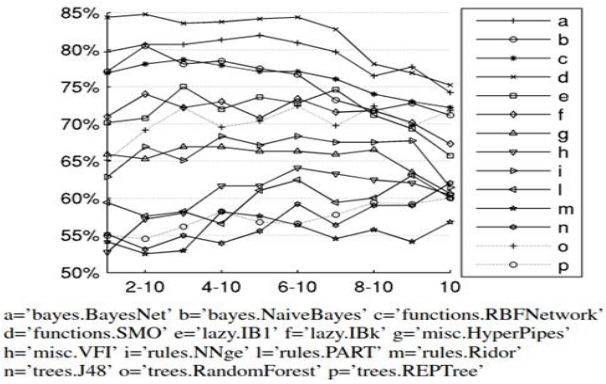


Figure 2: Examines the average percentage of accurate categorization for the seven emotional states inside the EMODB dataset. This analysis encompasses the use of diverse feature aggregates and classification systems.

The classifiers, including SMO, Simple Logistic, BayesNet, NaiveBayesSimple, NaiveBayes, and RBFNetwork, exhibit exceptional performance, with an average recognition rate ranging from 77% to 87%. On the other hand, the second set of classifiers, including IB1, Hyper Pipes, Random Forest, NNge, and KStar, demonstrate moderate levels of performance, as shown by an average recognition rate ranging from 67% to 77%. Last but not least, the recognition rates of the Part, J48, REPTree, Ridor, and VFI classifiers, which make up the third group, range from the lowest (55%) to the highest (66%). A

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

tendency of decreasing inequality can be seen between the three groups from the sixth to the tenth aggregate. One possible explanation for the trend is that the classifiers in the first group are becoming less effective while those in the third group becoming more so. Classifiers from the second category seem to be somewhat insensitive to the specific set of parameters used, at least in a broad sense. No matter what performance levels are measured using the subset 910, the SMO method always shows improvement. The SMO system has an identification rate of almost 87% when using the 2-10 aggregate, which is a 6% increase over the case when using all characteristics. Table 1 presents a matrix illustrating the misclassification of emotional states in the context of the specific event under consideration. Figure 3 depicts the performance of several categorization algorithms in terms of the average recognition percentage for the Activation, Evaluation groups and Eminem. This assessment is performed with various subsets of parameters. The lack of correlation between the performance and the aggregate parameter used is apparent. The Support Vector Machine (SVM) classifier demonstrates a higher level of performance in the majority of the instances that have been investigated. The evaluation of the entity being analyzed.

Table 1: In this study, we use the 2-10 aggregate technique and SMO classifier to assess the misclassification rate in distinguishing among seven distinct emotional states

	Neutral	Melancho	Happiness	Anxiety	Repugnanc	Weariness	Anger
Neutral	**88.46**	1.28	0.00	1.28	0.00	7.69	1.28
Melancholy	1.92	**98.08**	0.00	0.00	0.00	0.00	0.0
Happiness	0.00	0.00	**64.06**	6.25	1.56	0.00	28.12
Anxiety	3.64	0.00	7.27	**78.18**	0.00	1.82	9.09
Repugnance	0.00	2.63	0.00	7.89	**86.84**	2.63	0.0
Weariness	5.06	0.00	0.00	0.00	2.53	**92.41**	0.0
Anger	0.00	0.00	11.81	3.15	0.00	0.00	**85.04**

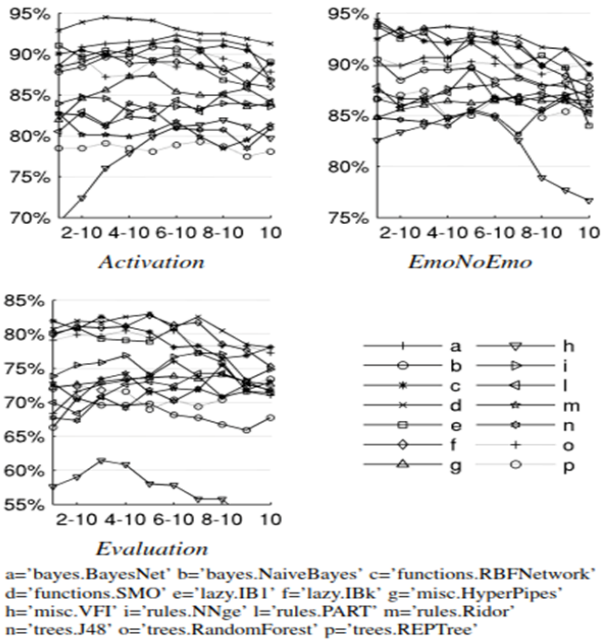


Figure 3: The average percentage of accurate categorization for the Activation, EmoNoEmo, and Evaluation groups was determined by using different parameter aggregates.

Table 2: By using the 3-10 aggregate and SMO classifier, the misclassification rate for the Activation subgroup may be computed.

	Neutral	Happiness-Fear- Repugnance-Anger	Melancholy-Weariness
Neutral	**87.18**	3.05	1.41
Happiness-Fear- Repugnance-Anger	5.34	**97.89**	5.13
Melancholy- Weariness	0.70	7.69	**91.60**

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

Table 3. The misclassification rate in the Eminem group was obtained using the 1-10 aggregate and the SMO classifier

	Neutral	Melancholy-Happiness- Fear-Repugnance- Weariness-Anger
Neutral	**76.92**	23.08
Melancholy-Happiness- Fear-Repugnance- Weariness-Anger	2.41	**97.59**

In the EmoNoEmo and Activation groups, accuracy in classifying emotional states exceeds 93%, whereas it drops to a maximum of 87% in the Evaluation group.

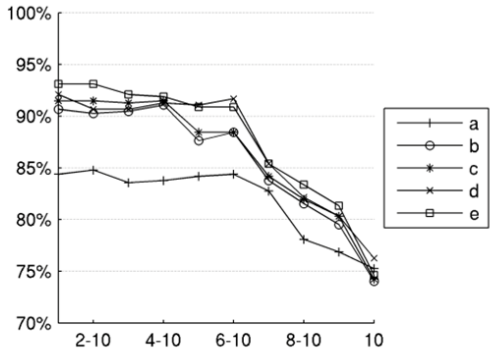
Table 2 displays the proportion of false positives and false negatives for each Activation cluster using the SMO classifier on the 2-to-10 aggregate. An impressive 95% classification accuracy is reached here, making this the best-case scenario for average performance. The classification of the Neutral condition has the worst performance. The words within this particular group are erroneously categorized as members of the Anger-Repugnance-Fear-Happiness group in more than 8% of instances, and as members of the Weariness- Melancholy group in over 4.5% of instances.

In the comparison of two groups, namely Emonoemo and SMO, it is seen that SMO once again demonstrates superior performance, as shown by the 1-10 aggregate measure. The mean performance achieved in this instance is 95.4%. However, a significant amount of misclassification is seen in the categorization of words associated with the Neutral state.

Based on the data shown in Table 4, it is apparent that a considerable percentage, over 22%, of instances are erroneously classified under the Melancholy-Happiness- Fear-Repugnance-Weariness-Anger category. In the pursuit of determining the most ideal mean value, the use of the collective known as Evaluation is employed.

Table 4. We determined the percentage of incorrect classifications made by the test group using the 5-10 aggregate and the SMO classifier.

	Neutral	Melancholy-Fear- Repugnance- Weariness-Anger	Happiness
Neutral	**71.79**	28.21	0.00
Melancholy-Fear-Repugnance- Weariness-Anger	1.56	67.19	**31.25**
Happiness	1.99	**94.87**	3.13



a='no discretization and normalization', b='Fayyad & Irani', c='Fayyad & Irani + normalization', d='Kononenko', e='Kononenko + normalization'

Figure 4: Using feature aggregation, SMO, discretization, and normalization, the typical success rate for labeling each of the EMODB's seven emotions.

We aggregate 5-10 parameters and utilize the SMO classifier to obtain the desired performance. Based on the data shown in Table 4, it seems that Happiness is incorrectly categorized as one of the Emotions of Anger, Repugnance, Fear, and Melancholy in over 68% of all cases. It's also important to note that the Neutral category is often misclassified, with a rate higher than 29%. Many people make this mistake because they incorrectly group Neutral with the emotions of anger, repugnance, fear, and melancholy.

Efforts were made to improve the classification system's efficiency after the initial study was completed, with discretization of features and normalization of scores being implemented as discussed in Sections 4.2 and 4.3. The data shown in Figure 4 illustrate that the Kononenko discretization technique and z-score normalization provide the highest performance. The average rate of correct identification, based on the assessment of parameters 1-10, is close to 94%. When compared to a perfect situation in which no discretization or normalization procedures are used, this implies an improvement of about 8 percent. In Table 5, we have a matrix depicting the incorrect labeling of feelings in this case. In Figure 5, we see a comparison of the typical proportion of correct categorization across different categories of mood. This comparison employs many aggregate characteristics, the SMO algorithm, as well as discretization and normalizing approaches. The use of approach and z-core normalization techniques yields optimal outcomes when implemented on the Activation and EmoNoEmo cohorts. When all attributes 1–10 are considered

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

together, we often get a classification accuracy of 98%–99%. Kononenko discretization with z- score normalization yields best results for the Evaluation subgroup. Combining characteristics 3-10, we find that the median share of

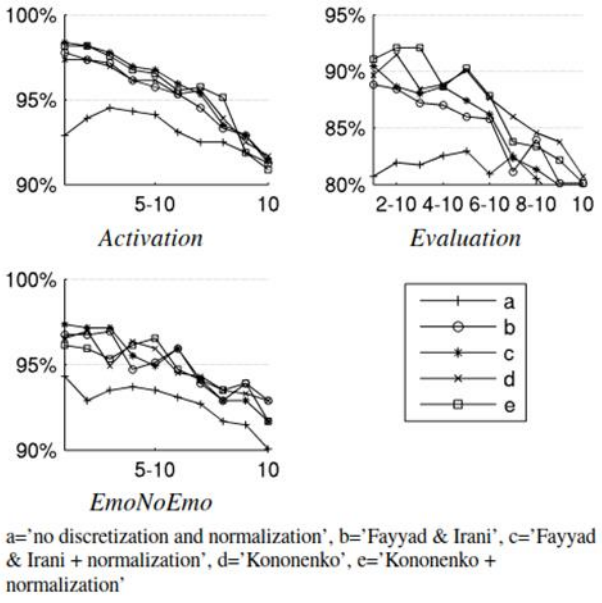


Figure 5: The mean accuracy rates for the Activation, EmoNoEmo, and Evaluation categories were calculated via the use of several feature aggregates, the SMO algorithm, and methods such as discretization and normalization.

The categorization has a level of accuracy about equal to 94%. Tables 6, 7, and 8 provide the misclassification rates according to emotional states in the previously discussed situations. Using the SUSAS speech corpus, the accuracy of the method used for the categorization of emotional states was verified. Differences between the dataset under consideration and EMO-DB are primarily based on three factors: language (Spanish versus American English), time scale (individual words vs. entire sentences), and the specific emotions represented. In accordance with the technique described in the previously stated work, we used a subset of 6 states from the SUSAS dataset, rather than using all 11 states that were provided. The analysis focused on evaluating the effectiveness of the SMO classification approach across nine distinct groups, namely Boston1, Boston2, Boston3, General1, General2, General3, New York1, New York2, and New York3. At the outset, the study encompassed all 3809 attributes obtained for each term. Then, the 10 groups of parameters were generated using the feature selection approach detailed in Section 4.1. Then, we implemented the discretization and normalization processes, as outlined in Sections 4.2 and 4.3, respectively. When just normalized and discretized characteristics were employed, the average recognition percentage jumped from about 80% to over 91%, showing a considerable increase. Figure 6 shows an example set of findings generated by combining the groups Boston1, General1, and NewYork1.

Table 5. We used the SMO classifier, the 1-10 aggregation approach, Kononenko discretization, and feature normalization to determine the misclassification rates between the seven distinct emotional states.

	Neutral	Melancholy	Happiness	Anxiety	Repugnance	Weariness	Anger
Neutral	**91.02**	1.28	0.00	1.28	0.00	5.12	1.28
Melancholy	1.92	**98.08**	0.00	0.00	0.00	0.00	0.0
Happiness	0.00	0.00	**81.25**	0.00	0.00	0.00	18.75
Anxiety	0.00	1.82	1.82	**90.90**	1.82	0.00	3.63
Repugnance	0.00	2.63	0.00	0.00	**94.74**	0.00	2.63
Weariness	2.53	0.00	0.00	0.00	0.00	**97.47**	0.00
Anger	0.00	0.00	2.37	1.57	0.00	0.00	**96.06**

Table 6. The 1-10 aggregate and the SMO classifier and method were used to determine the misclassification rates for the Activation subgroup.

	Neutral	Happiness-Fear-Repugnance- Anger	Melancholy-Weariness
Neutral	**96.16**	2.56	1.28
Happiness-Fear-Repugnance- Anger	1.53	0.00	**98.47**
Melancholy-Weariness	0.70	**98.94**	0.36

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

Table 7. The percentage of misclassification in the EmoNoEmo group was determined using the 1-10 normalization techniques and the SMO classifier for feature processing.

	Neutral	Melancholy-Happiness- Fear-Repugnance- Weariness-Anger
Neutral	**93.59**	6.41
Melancholy-Happiness- Fear-Repugnance- Weariness-Anger	1.93	**98.07**

Table 8. By using the 3-10 aggregate methodology and the SMO classifier, the Kononenko discretization and normalization methods on the dataset, we successfully computed the misclassification percentage within the Evaluation group.

	Neutral	Melancholy-Fear-Repugnance- Weariness-Anger	Happiness
Neutral	**89.75**	10.25	0.00
Melancholy-Fear-Repugnance- Weariness-Anger	0.0	21.88	**78.12**
Happiness	1.70	**95.16**	3.14

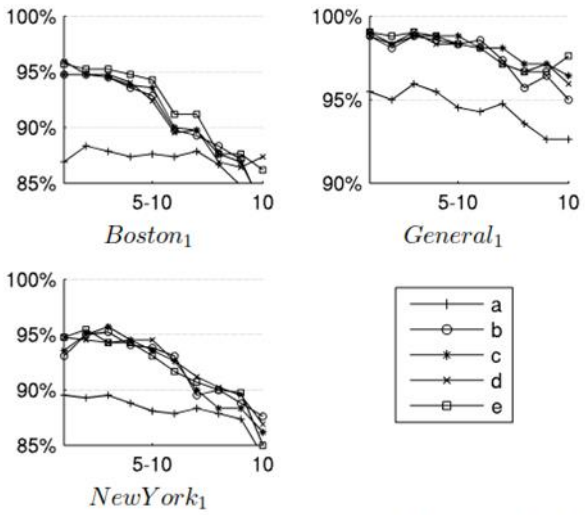


Figure 6: Using a variety of feature aggregation, SMO algorithm, discretization, and normalization approaches, we calculated the average accuracy rate for accurately categorizing the six emotional states in the SUSAS DB.

CONCLUSION

This paper examines the implementation of an autonomous system for detecting emotional states inside a DSR (Dynamic Systems and Robotics) framework. The system uses a conventional front-end and characteristics derived from an audio stream. Both the Spanish and the American English speech corpora were used in the study. More than 3800 statistical components were employed to create a feature vector from each audio segment. The pitch period, jitter, and voicing class in addition to the 13 cepstral coefficients used in the energy temporal trend approach were used to produce the aforementioned elements. A comparative analysis was performed on several machine learning techniques offered by the KDDM software. The superiority of the SMO algorithm in terms of performance has been shown. The use of a correlation-based methodology for the selection of features, along with subsequent normalization and discretization of the selected features, yielded a noteworthy improvement in performance. The classification accuracy attained with the EMO-DB dataset surpassed 93%. Moreover, the system exhibited a reasonable level of efficiency in categorizing the nine occurrences of dialect in the SUSAS dataset. All accuracy percentages beyond 93% are observed, with the notable exception of GENERAL3, which attains a flawless accuracy rate of 100%. This implies that the classification of all 425 recordings was executed with meticulousness, guaranteeing their accurate identification as pertaining to the six emotional states under investigation.

REFERENCES

[1] Al-Dujaili, M. J., & Ebrahimi-Moghadam, A. (2023). Speech emotion recognition: a comprehensive survey. Wireless Personal Communications, 129(4), 2525-2561.
[2] Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, 245-263.

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq

[3] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.
[4] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345.
[5] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. IEEE access, 9, 47795-47814.
[6] Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 10(10), 1163.
[7] Bao, F., Neumann, M., & Vu, N. T. (2019, September). CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition. In Interspeech (pp. 2828-2832).
[8] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), 90-99.
[9] Liu, Z. T., Xie, Q., Wu, M., Cao, W. H., Mei, Y., & Mao, J. W. (2018). Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing, 309, 145-156.
[10] Mustaqeem, & Kwon, S. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, 20(1), 183.
[11] Hassan, M. D., Nasret, A. N., Baker, M. R., & Mahmood, Z. S. (2021). Enhancement automatic speech recognition by deep neural networks. Periodicals of Engineering and Natural Sciences, 9(4), 921-927.
[12] Kadhim, I. B., Khaleel, M. F., Mahmood, Z. S., & Coran, A. N. N. (2022, August). Reinforcement Learning for Speech Recognition using Recurrent Neural Networks. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5). IEEE.
[13] Laylani, L. A. A. S. S., Coran, A. N. N., & Mahmood, Z. S. (2022, January). Foretelling Diabetic Disease Using a Machine Learning Algorithms. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-5). IEEE.
[14] Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. IEEE Transactions on affective computing, 6(1), 69-75.
[15] Ghosh, S., Laksana, E., Morency, L. P., & Scherer, S. (2016, September). Representation learning for speech emotion recognition. In Interspeech (pp. 3603-3607).

*Corresponding author
Ali Najdet Nasret,
Electrical Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
e-mail: alinajdet@ntu.edu.iq