

## Research Article

# A Summary of Work on Emotional Voice Recognition with Machine Learning

Manroop Kaur

UG Student (Department of Computer Science and Technology), DAV Institute of Engineering and Technology, Jalandhar, Punjab, India.

## I N F O

**Corresponding Author:**

Manroop Kaur, (Department of Computer Science and Technology), DAV Institute of Engineering and Technology, Jalandhar, Punjab, India.

**E-mail Id:**

manroop2107@gmail.com

**Orcid Id:**

<https://orcid.org/0000-0003-4970-3412>

**How to cite this article:**

Kaur M. A Summary of Work on Emotional Voice Recognition with Machine Learning. *J Adv Res Comp Graph Multim Tech.* 2023; 5(1): 1-13.

Date of Submission: 2023-02-18

Date of Acceptance: 2023-05-01

## A B S T R A C T

This chapter compares SER systems. This talk covers a theoretical definition, affective state categorization, and several techniques to express emotions. This inquiry requires an SER system with many classifiers and feature extraction methods. After extracting Mel-Frequency Cestrum Coefficients (MFCC) and Modulation Spectral (MS) properties from speech signals, they are utilized to train multiple classifiers to classify them. FS was used to find the most relevant feature subset. Emotion classification was solved using many machine learning models. The initial categorization of these seven sentiments uses an RNN classifier. After that, their results are compared to those from spoken audio signals emotion detection methods like Multivariate Linear Regression (MLR) and Support Vector Machine (SVM). Experimental data has been collected from Berlin and Spanish databases. Speaker Normalization (SN) and feature selection improve the accuracy of all Berlin database classifiers to 83%. In Spanish datasets, RNN classifiers without SN and with FS have the maximum accuracy (94%).

**Keywords:** Mel-Frequency Cestrum Coefficient (MFCC), Support Vector Machines (SVM), Modulation Spectral Features (MSFs), Recursive Feature Elimination (RFE), Multivariate Linear Regression Classification (MLR), SVM, and RNN are used in this Research

## Introduction

Emotions express a person's unique view of a thing, event, or condition via consciousness, physical feeling, and behavior. Happiness, sorrow, rage, contempt, and everything in between are emotions. Human relationships depend on emotion. This helps make smart and sensible decisions. When we share our thoughts and get criticism, we can relate to others. Research shows that emotion affects people's interactions. The "wheel of emotions" was created by Robert Plutchik in the 1980s. This showed how many emotions may be mingled. Emotion recognition detects human emotions. How someone expresses their emotions may reveal their mental health. Because of this, automated

emotion identification, a new branch of study, aims to analyses and remember happy experiences. Understanding human speech emotions has grown increasingly crucial to better human-machine interactions. Due of AI's growth. Finding emotions might be difficult for numerous reasons. Defining feeling is tricky. It's tough to link signal patterns to emotions. Emotion-relevant communication patterns vary from person to person and context to circumstance.

Previous research has examined facial expressions, voice, physiological data, and more to discern emotional states. Speech signals' benefits make them suited for affective computing. Affective computing benefits from speech. Voice transmissions are simpler and cheaper than many other

biological signals. This draws most researchers to Speech Emotion Recognition (SER). SER analyses the speaker's emotions from their words. The scientific community's interest has grown in recent years. The interface between humans and machines, audio surveillance, web-based e-learning, commercial applications, clinical research, entertainment, banking, contact centres, cardboard systems, computer games, and more can use the ability to read emotions.

The SER system must identify an appropriate emotional speech database, extract useful characteristics, and utilise machine learning to create reliable classifiers. The SER system's biggest issue is identifying emotions. Two further ways to show an individual's emotions are:

- distinct • Dimensional Representation is the process of representing feelings using dimensions like valence (from negative to positive), activation or energy (from low to high), and dominance (from active to passive)

Both methods have pros and cons. The dimensional technique is more sophisticated and provides more prediction data, but it is harder to implement since there is less annotated audio data in dimensional formats. This method's limitation. Dimensional representation provides more context for predictions than discrete categorization, which is simpler to understand and implement.

LPCC, MFCC, and MSFs may convey emotional information in speech, according to many experts.<sup>4</sup> Extraction and selection of features may increase learning, processing difficulty, model generalizability, and storage space. Classification completes emotion recognition in spoken speech. It involves categorizing raw data in the form of a speech or frame into an emotion based on inferred features. Data has these characteristics.

Speech emotion recognition researchers have suggested several categorization techniques recently. The Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), Neural Networks (NN), and Recurrent Neural Networks (RNN) are among these approaches. Other speech-emotion classifiers include a modified Brain Emotional Learning model (BEL).<sup>19</sup> MLP and ANFIS are combined in this model.

Sections of the chapter follow. Next, a quick literature review. In the third section, we will examine speech emotion basics. Next, we'll explore speech signal properties. We provide SER feature selection and machine learning algorithms. Our datasets and simulation results using different characteristics and Machine Learning (ML) paradigms are presented in the following section. Section 6 presents the experiment results. Sections 7 and 8 conclude this chapter with conclusions and ongoing work.

## Literature Evaluation

In recent years, speech emotion detection has increasingly depended on a range of machine learning algorithms, each of which considers a distinct collection of variables.

A. Milton, S. Sharmy Roy, and S. Tamil Selvi, PhD employed a three-stage Support Vector Machine classifier to classify the seven sensations in the Berlin Emotional Database.<sup>1</sup> The classification procedure uses all 535 database files to extract MFCC traits. These phrases frames undergo nine statistical measures. Hierarchical SVM uses linear and RBF kernels and sets the RBF sigma to one. Ankur Sapra, Nikhil Panwar, and Sohan Panwar<sup>2</sup> have suggested using pitch, energy, and other sonic cues to determine human emotions. The suggested system classifies data using the closest neighbor algorithm and the conventional MFCC approach.<sup>2</sup> Because male and female voices have different tone ranges,<sup>1,4</sup> their MFCCs have different emotions. MingyuYou, ChunChen, and JiajunBu<sup>4</sup> suggested a clear and loud speech-based emotion recognition system. Geodesic distance was needed to preserve emotional communication geometry. An enhanced Lipschitz embedding was utilised to embed 64-dimensional audio characteristics into a six-dimensional space using the geodesic distance estimation. Direct emotion identification from loud speech avoided noise reduction issues. Daniel Neiberg, Kjell Elenius, Inger Karlsson, and Kornel Laskowski<sup>8</sup> used Mel-Frequency Cepstral Coefficients (MFCCs) to forecast pitch between 20 and 300 Hz. Mel-frequency Cepstral Coefficients estimated these. Basic pitch properties are also presented. GMMs have been used entirely to represent these frame-level acoustic properties. Kwang-Dong Jang and Oh-Wook Kwon<sup>7</sup> tested a voice emotion identification strategy during emotional human-robot communication. Shock, wrath, boredom, delighted, neutral, and pleasure are the recommended categories. After removing background noise and identifying spoken words, we compile phonetic and prosodic data to construct a feature vector for each speech. Phonetic information comprises log energy, shimmer, formant frequencies, and Teaser energy. Prosodic information includes pitch, jitter, and speech speed. Rhythmic information is prosodic. After that, a pattern classifier using Gaussian support vector machines classifies the speech by mood.

A summary of your database corpuse categories follows.

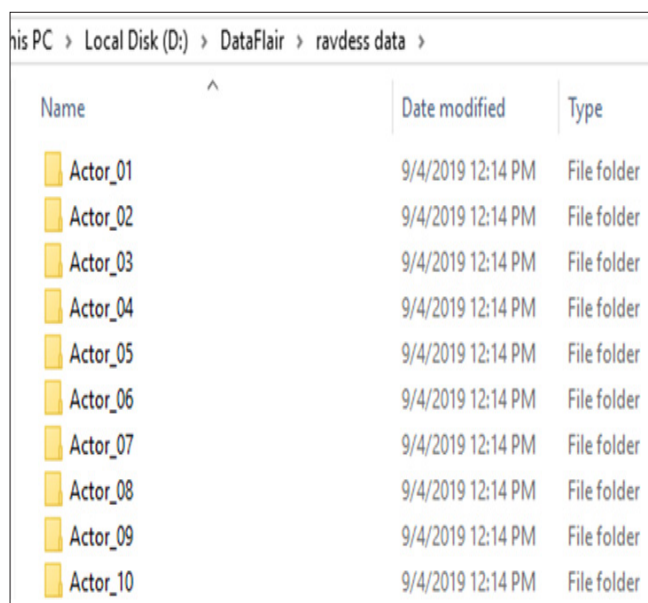
## Databases-Review

High-quality emotional speech is essential for emotion characterization, whether for synthesis or recognition. Speech corpora have three categories.<sup>12</sup> These corpora study is summarised here.

## Emotion Database From Actors

Radio artists with experience and training may contribute

to actor-based or simulation-based emotional speech corpora. “Full-blown feelings” may also describe some emotions. A speaker-independent Tamil database,<sup>10</sup> Burmese and Mandarin databases,<sup>17</sup> and the Danish Emotional Speech Database,<sup>18</sup> recorded with four radio performers, two male and two female-are popular. The IITKGP-SESC<sup>13</sup> was recorded by professional artists from All India Radio, Vijayawada, while the Emotional Speech Database for Basque<sup>9</sup> has professional dubbing artists recording six key emotions. Both are corpora. The SHURE dynamic cardioid microphone recorded 16 kHz speech samples. 12000 statements are made. Thus, each of the eight emotions-anger, disgust, fear, happiness, neutral, surprise, sarcasm, and compassion-has 1500 phrases. The RODE-NT2 condenser cardioid microphone helped<sup>1</sup> create a database of emotional speech. Ten speakers with neck impedances were laryngography simultaneously.



Name	Date modified	Type
Actor_01	9/4/2019 12:14 PM	File folder
Actor_02	9/4/2019 12:14 PM	File folder
Actor_03	9/4/2019 12:14 PM	File folder
Actor_04	9/4/2019 12:14 PM	File folder
Actor_05	9/4/2019 12:14 PM	File folder
Actor_06	9/4/2019 12:14 PM	File folder
Actor_07	9/4/2019 12:14 PM	File folder
Actor_08	9/4/2019 12:14 PM	File folder
Actor_09	9/4/2019 12:14 PM	File folder
Actor_10	9/4/2019 12:14 PM	File folder

**Figure 1. Burmese and Mandarin Databases<sup>12</sup>**

### Elicited Emotional Speech Database

To create elicited speech corpora, an artificial emotional environment is replicated in a manner that is imperceptible to the speaker. Induced speech corpora is another name for these data sets. There is a lot of labour involved in classifying the database in this way. This has resulted in a significant reduction in the number of publicly available evoked speech corpora. If recognition systems are not adequately trained using the right database, their performance and resilience will suffer almost immediately. To successfully train and subsequently assess the efficacy of an emotion recognition system, a large enough corpus of relevant words is required. There are three main categories of emotional registries: those including enacted emotions, those containing emotions that occur naturally, and those containing emotions that are prompted.<sup>2,4</sup> In the past, we

have consulted emotion databases, which include robust examples of numerous emotions, for use in our work. Most research on speech emotion recognition<sup>4</sup> has used speech recorded when the speaker was trying to convey an emotion. Each the Berlin Database and the Spanish Database were utilized in our tests to categories different emotions into distinct groups; this section gives a detailed description of each of these resources.

### Collection of Mood and Feeling Expressions from the Wild

Call center transcripts, cockpit recordings, and people's emotional reactions to one another in public areas are all examples of data collected in its natural habitats. That's why we know they're the real deal. The German television chat show "Vera am Mittag" was the source of the database utilised in.<sup>15</sup> Differentiating between the recordings required adjusting the volume of each speaker individually. We use the three most fundamental emotions-also known as emotion primitives-to describe the range of feelings in an emotion space. They are real in the valence, activity, and dominance configurations that best suit them.

### Emotions and Categorization

Here, we'll take a look at what "emotion" really means and at the many hypotheses that have been put up to explain it. There are several methods and sources of information that may be used to successfully determine emotional states. Here, we'll break down each process into its essential steps.

### An analysis of the Emotion

Emotions are mental states that may be categorized as either "subjective," "physical," or "behavioral." An emotional response is a deliberate mental reaction characterized by a strong feeling of attachment to a particular object and manifested in a variety of physiological and behavioral changes in the organism. Emotions are experienced in response to external stimuli.

Attempting to pin down the essence of human feeling is a daunting undertaking for psychologists. The idea of emotions, in reality, has a number of different interpretations in the scientific literature. An individual's emotional state is intricately intertwined with their general disposition, demeanor, and character. Many psychologists agree that emotions are "complicated conditions of feeling" that cause physiological and cognitive changes in the individual experiencing them. People's attitudes and actions are influenced by these modifications. Others argue that emotions are better understood as syndromes comprised of factors including motivation, mood, behavior, and physiological changes, rather than as actual causal factors. What an emotion is and what it does might be understood in many ways depending on who you ask. It might be argued that emotions are physiological states

triggered by mental, emotional, and behavioral processes. A wide range of emotions, from elation to dishonesty, is also possible. However, there is little consensus among emotion theorists on the core elements of the emotion idea.

However, it seems that the great majority of people share the view that emotions are useful. For instance, they prepare us to react to both good and negative external stimuli (like being rejected by an interviewer). We can feel fear in this position, and that might make us want to retreat. A person's disposition, feelings, affects, and sense of well-being may all be categorized under the umbrella word "emotion".<sup>26</sup> According to,<sup>6</sup> "emotion" may also refer to a state that is very nuanced and intertwined with several cognitive, physiological, and somatic mechanisms.

### Different Types of Emotions

Researchers have classified emotions in a wide variety of ways, but the most consistently observed core emotions in the great majority of studies are joy, sadness, anger, fear, disgust, and surprise. The valence-arousal plane<sup>18</sup> is a two-dimensional representation of these emotions. Joy and surprise are examples of positive emotions, whereas wrath, fear, and disgust are examples of negative emotions.<sup>12</sup> Two instances of pleasant feelings are joy and surprise. Prolonged exposure to adverse circumstances has the ability to increase a person's emotional stress.<sup>7</sup> Happiness, sadness, wrath, fear, disdain, and surprise are all examples of primary emotions, whereas the construction of a mental image that corresponds to a memory or a core feeling is an example of a secondary emotion.<sup>8</sup>

It's possible to categories feelings on a wide variety of scales.<sup>12</sup> Arousal and valence provide the foundation for these scales and dimensions. In contrast to the binary nature of the valence dimension, the arousal level may range from "not-aroused" to "excited." For instance, shock causes a high amount of arousal, whereas disgust and grief both cause low arousal.<sup>8</sup> The emotional value of surprise is similar to that of delight, however the valence of distaste is negative. The valence and arousal-based four-class paradigm is shown graphically in Figure 2.

### Discourse

Some people's voices are even more important than their faces when it comes to portraying their inner states of mind. Since the speaker's emotional state may be inferred from the tone of their voice, speech is a powerful medium of communication that is enriched by sentiments. Several key voice feature vectors will be investigated in this study. These include fundamental frequency, Mel-Frequency Cepstral Coefficient (MFCC), Prediction Cepstral Coefficient (LPCC), and others.

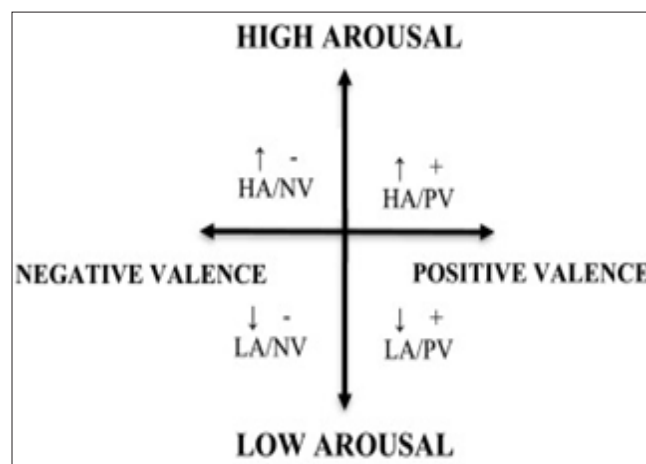


Figure 2. Illustration of Four Class Models based on Valence and Arousal<sup>3</sup>

### Symptoms and Indications that Occur in the Body

The autonomic nervous system's ability to send forth physiological information allows for an impartial examination of emotions. These include measurements of the Skin's Temperature (ST), Respiratory Rate (RSP), Blood Pressure (BP), and Heart Rate (HR).<sup>3</sup> Skin Temperature (ST), Blood Volume Pulse (BVP), and Skin Conductance (SC) are some other measures. Those who struggle to put their feelings into words might benefit from the use of physiological signs to discern emotions. It's possible that these people's mental or physical health prohibits them from expressing emotion.

The human face is very expressive and can convey a wide range of feelings even when words aren't needed to do so.<sup>1</sup> In contrast to certain other forms of nonverbal communication, facial expressions are rather widespread. Overwhelmingly, people of different cultures utilise the same facial expressions to represent different emotions.

The Proposed System for Automatic Speech and Emotion Recognition (SER).

The term "Speech Emotion Recognition" (or "SER" for short) describes the method of deducing a person's emotional state from their verbal output alone. Using a large database of sound clips, the system can identify the emotions being expressed. This method involves listening for patterns in a person's continuous and unprompted speech to determine how they are feeling. The six components of this system are as follows: vocal activity detection; speech segmentation; signal pre-processing; feature extraction; emotion classification; and frequency analysis. There are typically four stages to the SER procedure. Recording some of your own voice is the first order of business.

After the first set of features has been retrieved, the second

set of features vectors is created based on the emotions associated with the speech.

Following this, we attempted speech emotion feature selection, or identifying the characteristics that should be utilized to distinguish between the various emotions.

A machine learning classifier, also known as speech emotion classification recognition, is trained with these properties so that it may be utilised for recognition. Figure 3 depicts this method in detail.

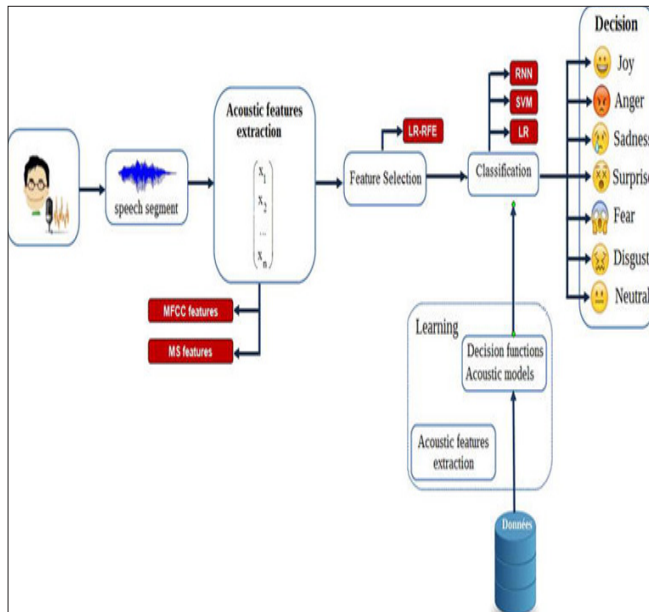


Figure 3. Block Diagram of the Proposed System<sup>13</sup>

### Collection of Data

“Big data” refers to gathering, combining, and storing huge databases for subsequent study. This strategy is “big data.” Depending on the collection, this data might be sorted, left unsorted, or mixed. Order and disorder are possible. The same process records sound. Audio samples and data demand a lot of space. For your convenience, the following formats may save audio on digital devices:

The digital audio coding standard “MPEG-1 Audio Layer 3” is reduced to “MP3,” a frequent abbreviation.

Computers can store audio bitstreams as wav files. It stands for “Waveform Audio File,” the file’s format.

Microsoft’s Windows Media Audio (WMA) standard includes codecs and audio coding types.

Voice-activated software improves with more multilingual audio data. The audio data gathering section details the process of collecting and analysing audio data from various sources. ASR systems and virtual assistants require massive audio data to train their speech recognition algorithms. Information collecting strategies:

**Increase Information:** The data is recorded with standardised noise indicators, the sequence of measurements, a noise description, and other valuable information.

**Reliable Data Collection:** Internal stress waves, structural vibrations, and structure-fluid interactions through acoustic radiation all provide information.

**Natural Language Utterance Collection** components include: This is a rich trove of natural language data from user profiles and other websites.

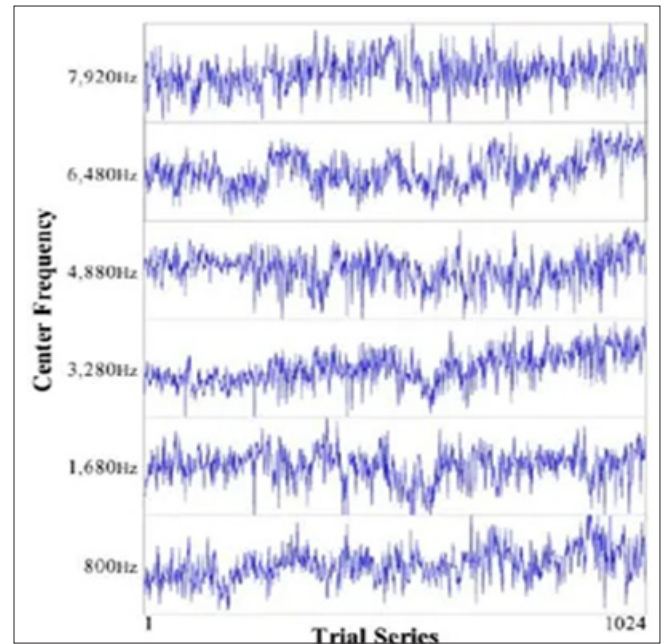


Figure 4. Frequency of Various Audio Files<sup>12</sup>

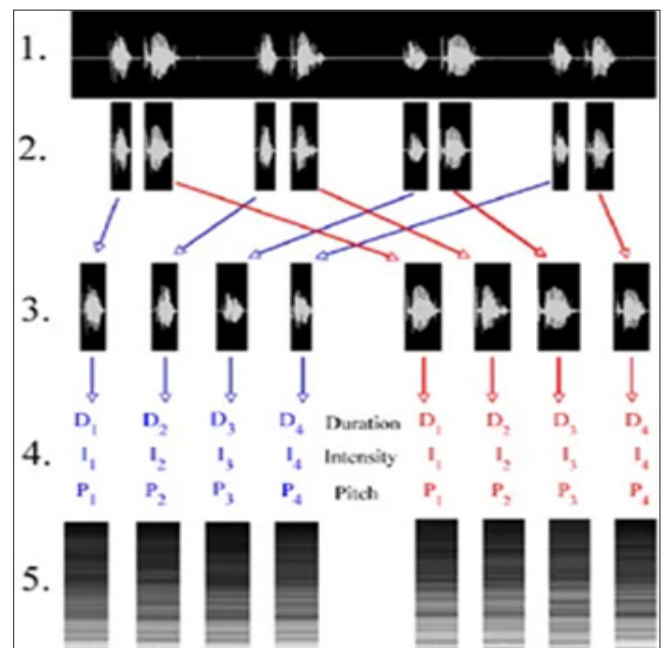


Figure 5. Steps in Data Collection<sup>12</sup>

stages In the Collection of data the following are the stages that are involved in the collection of audio data:

1. Remove individual utterances from audio recordings and save them as blocks.
2. The audio files are segmented using a syllable-based system
3. The audio recordings are then sorted in a sequential fashion according to the syllables.
4. It is necessary to establish the values for the strength, duration, and pitch of each syllable.
5. Plots of the estimated spectral power are shown in a way similar to a spectrograph.

### Berlin Emotional Database

At the Technical University of Berlin, 10 professional artists-five men and five women-recorded the Berlin Emotional Database.<sup>7</sup> This speech corpus is acknowledged internationally. The multi-speaker database has 7 emotions and 535 German expressions. Speaker-independent testing is simple with it. For database development, it has five short sentences and five long phrases, with an average of 1.5 to 4. Twenty people evaluated 800 raw database utterances. The database was recorded using a 16-kHz Sennheiser MKH 40 P48 microphone.<sup>16</sup> 16-bit numbers store samples. Leave-one-out cross-validation testing is easy with this database.<sup>15</sup> Table 1 lists the states of consciousness and the quantity of speech files in the database.

**Table 1. Number of Emotional Speech Files in Berlin Database<sup>3</sup>**

Emotion	No. of Speech Files
Anger (A)	127
Boredom (B)	81
Happiness (H)	71
Fear (F)	69
Sadness (S)	62
Disgust (D)	46
Neutral (N)	79
Total Number of Files	535

### Spanish Database

Spanish emotional emotions include words from male and female professional actors. The “six basic emotions plus neutral” (anger, grief, joy, fear, disgust, surprise, and neutral/normal) were each referenced twice in our Spanish corpus (which may be utilised for free for academic and research reasons).<sup>4</sup> Once the remaining four neutral variations-soft, loud, slow, and fast-were captured. This database is better since academics may use it and it includes 6041 utterances. This post focused on seven key emotions

from the Spanish database to compare with the Berlin database and improve recognition rates. This was done to increase recognition.

### Extracting Features

Pay attention to the voice signal’s many emotional cues. Emotion identification’s hardest issue is which traits to utilise. Recent study has shown similar properties such energy, pitch, formant, Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum coefficients (MFCC), and modulation spectral features. These are some recovered traits. We extracted this study’s emotional aspects using modulation spectral features and the MFCC. 4.1.2.1. Mel-Frequency Cepstrum Coefficient (MFCC).

It expresses voice transmission spectral properties most typically. Speech recognition is greatest with them since they consider frequency sensitivity. The Mel-frequency scale was used to translate the Fourier transform and energy spectra for each frame. The classification approach used MFCC values from the Discrete Cosine Transform (DCT) on Mel log energies. DCT coefficients 1-12 produced these results. It simplifies calculations, enhances discrimination, and resists noise, among other benefits. The Praat software provides MFCC characteristics with 20- and 10-millisecond window durations.<sup>6</sup> The hamming window is preferred for its high frequency resolution and side lobe suppression. The zero-crossing rate eliminated the database’s solitary parts, and the energy was thresholded. This decreased data. The silent zone was eliminated since it provided little valuable information. The MFCC uses the Mel scale,<sup>14</sup> which has linear spacing below 1000 Hz and logarithmic spacing above 1000 Hz, since human hearing is not linear. Human hearing is non-linear, hence this scale is used. The algorithm below calculates Mel frequency from any Hz frequency.

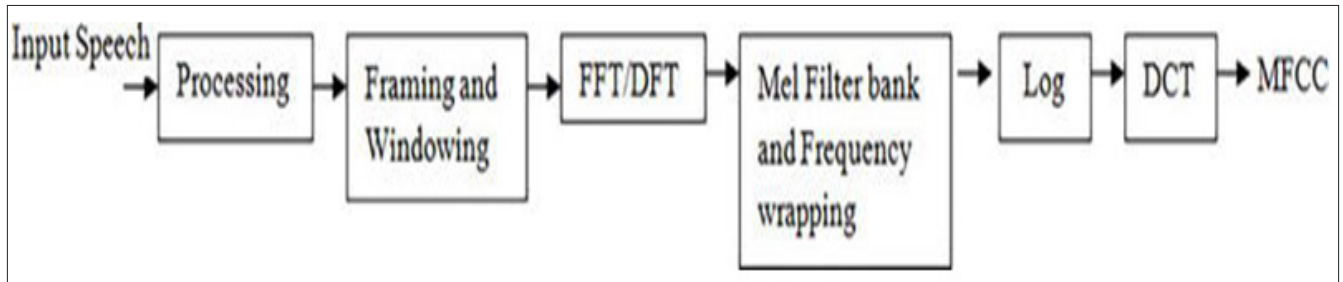
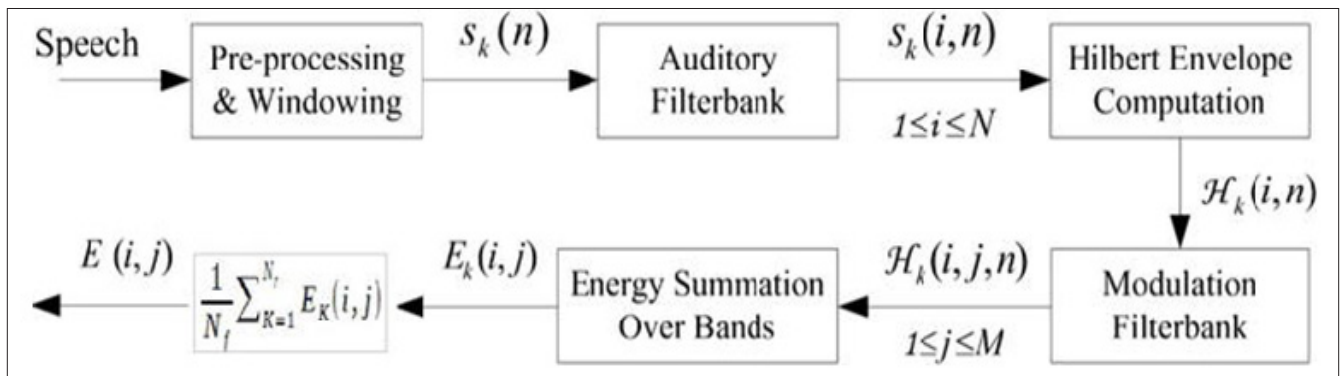
$$\text{Mel}(f) = 2595 \times \log\left(1 + \frac{f}{700}\right) \quad (1)$$

The Mel scale filter bank is a triangular array of equally overlapping filters with 50 centre frequencies and 100 constant bandwidths. Human auditory systems accomplish this.<sup>8</sup> Mel frequency scale spacing matches this.

Demonstrates MFCC computation.

### Modulation Spectral Features (MSFs)

A long-term spectro-temporal representation inspired by auditory signals is used to extract Modulation Spectral Features (MSFs). These characteristics are generated by simulating the human auditory system’s Spectro-Temporal (ST) processing, which considers both the regular acoustic frequency and the modulation frequency. Figure 7, shows the procedures for computing the ST representation.

Figure 6. Schema of MFCC Extraction<sup>13</sup>Figure 7. Process for Computing the ST Representation<sup>14</sup>

A 19-filter auditory filter bank segments spoken input to create the ST representation. Modulation signals are computed using critical-band output Hilbert envelopes. After that, a modulation filter bank processes Hilbert envelopes for frequency analysis. Modulation Spectral Features (MSFs) are the recommended modulation signal components.<sup>5</sup> The energy of the decomposed envelope signals is measured as a function of the modulation frequency and then the conventional acoustic frequency to obtain the ST representation. Energy averaged over all frames and spectral bands shows a feature. This experiment uses  $M=5$  modulation filters and  $N=19$  auditory filters. This study uses 95 (195) ST-derived MSFs.

### Choosing Features

Aha & Bankert<sup>34</sup> say feature selection in Machine Learning (ML) “reduces the number of features used to characterise a dataset so as to improve a learning algorithm’s performance on a given task.” ML feature selection seeks this. A specific learning strategy will be used to attain the maximum classification accuracy in a given assignment. The categorization model will need less features unintentionally. Feature Selection (FS) selects a subset of relevant features from the original features to improve recognition accuracy.<sup>5</sup> Relevance criteria underpin this strategy. It might significantly speed up learning algorithms. In this section, we will explain LR-RFE, an efficient feature selection approach employed in our research.

### RFE

Recursive Feature Elimination (RFE) uses a model like linear

regression or SVM to eliminate the feature with the greatest or worse performance. Recursive Feature Elimination (RFE) selects features by analysing smaller and smaller sets of features since these estimators weight features (for example, linear model coefficients). After training the estimator on the initial features, each feature’s prediction strength is determined.<sup>6</sup> Removing less important elements reduces the present set of attributes. After selecting the required number of features, the technique iterates recursively on the deleted features. We developed the LR-RFE feature ranking method using linear regression.<sup>7</sup> RFE research uses various linear models such the SVM-RFE, a feature selection approach based on the SVM. Guyon et al. chose key features using SVM-RFE. It may reduce categorization time and improve accuracy.

### Classification Methods

Several machine learning methods have been applied to categorise emotions. These algorithms use training examples to classify fresh data. Since every learning algorithm has pros and cons, there is no one right choice. Linear and nonlinear classifiers recognise a speaker’s emotions.<sup>2</sup>

### Nonlinear Classifiers

Non-linear classifiers use a weighted mixture of an object’s values, whereas linear classifiers use a linear combination of its properties.

### HMM

Hidden Markov For voice recognition, models are

preferred.<sup>17</sup> Speech applications have traditionally used HMM. A first-order Markov chain with disguised states protects the model's underlying structure. Model hidden states mirror data dynamics.<sup>2</sup> Left-to-right HMMs predominate.

K-Nearest Neighbour Classifier follows.

The K-Nearest Neighbour classifier, a simple machine learning method, can identify an object by its Euclidean neighbours. Large classes have more members when k is large. KNN fails if k is too low. K less than creates a classifier with less bias. The classifier has less bias.

### SVMs

Support vector machines need a kernel function to turn the original data set into a high-dimensional feature space.<sup>6</sup> This technique makes input samples linearly separable.<sup>14</sup> Figure 8 illustrates this with a full separation hyperplane. SVM classification performance is its biggest benefit, even with poor training data.

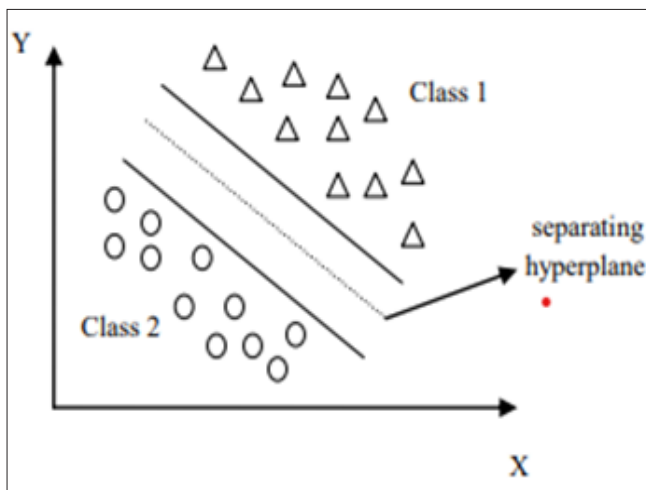


Figure 8.SVM Structure<sup>4</sup>

For linearly separable data points, classification is done by using the following formula<sup>3</sup>

$$\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1 \quad (2)$$

$$\langle w \cdot x \rangle + b_0 \leq -1, \forall y = -1 \quad (3)$$

where,  $(x, y)$  is the pair of training set. Here,  $n \times x \in \mathbb{R}$  and

$$y \in \{+1, -1\}.$$

$$\langle w \cdot x \rangle$$

represents the inner product of  $w$  and  $x$  whereas  $0$   $b$  refers to the bias condition.

SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function<sup>18</sup> is used here. The linear

$$(x, y) = (x \cdot y) \quad (4)$$

kernel function is given by the formula below,

$$\text{Kernel}(x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} \quad (5)$$

The radial basis kernel function is given by the following formula,

### MLR Classification

It optimizes resource utilisation and classification and regression problems using a simple machine learning algorithm computation. The LRC algorithm was tweaked as shown below. 1.<sup>9</sup> Step 3 determined the absolute value of the difference between the actual response vector and the predicted response vector ( $y_i$ ) rather than the Euclidean distance.

### RNNs are Next

It performs well in classification tasks and can learn time series data.<sup>42</sup> RNN models can learn temporal correlations, but vanishing gradient grows worse as the training sequence length rises. Hochreiter et al.<sup>4</sup> suggested using Long Short-Term Memory (LSTM) RNNs, which store data in memory cells to take advantage of long-range relationships.<sup>17</sup> Fig. shows a fundamental RNN implementation concept. The RNN uses the same values for its parameters ( $U$ ,  $V$ , and  $W$ ; see Figure for examples) throughout all layers, unlike traditional neural networks. Hidden state equations with variables:

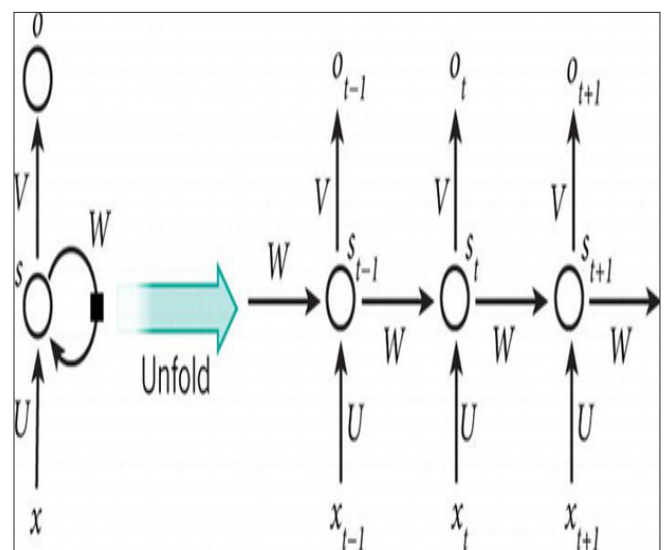


Fig 9, basic concept of RNN and unfolding in time of the computation involved in its forward computation<sup>13</sup>

## Data Exploration

The original sources' amalgamated data set is examined for the following:

- Emotional quickness and intensity differences
- Emotion-related energy levels

### Gender Emotional Distribution

The gender disparity was not large, however, there were somewhat more female speakers than male speakers.

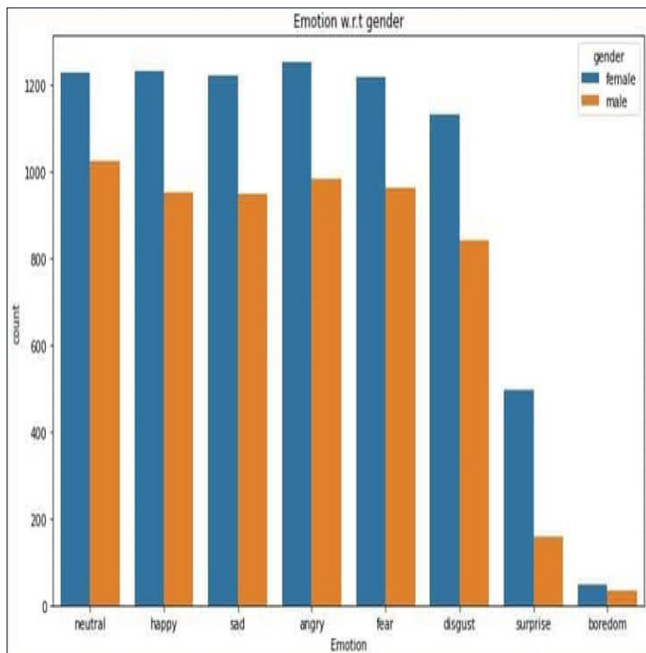


Figure 9. Distributions of Emotion with Respect to Gender<sup>12</sup>

### Variation in Energy Across Emotions

Because the audio samples in our dataset were of varying durations, we discovered that power, which is energy divided by unit time, was a more accurate way to assess energy fluctuation. This helped us maintain consistency in our investigation of energy variation. This metric was plotted based on the many feelings that were taken into consideration. It is fairly clear from looking at the graph that the major means by which individuals communicate their anger or fear is via the supply of a bigger amount of energy. In addition, we find that feelings of disgust and grief are closer to being neutral in terms of energy, despite the fact that there are outliers.

### Variation of Relative Pace and Power with respect to Emotions

A scatter-plot of power vs relative pace of the audio clips was analyzed and it was observed that the 'disgust' emotion was skewed towards the low pace side while the 'surprise' emotion was skewed more towards the higher pace side. As

mentioned before, anger and fear occupy the high-power space and sadness and neutral occupy the low power space while being scattered pace-wise. Only, the RAVDESS dataset was used for plotting here because it contains only two sentences of equal length spoken in different emotions, so the lexical features don't vary and the relative pace can be reliably calculated.

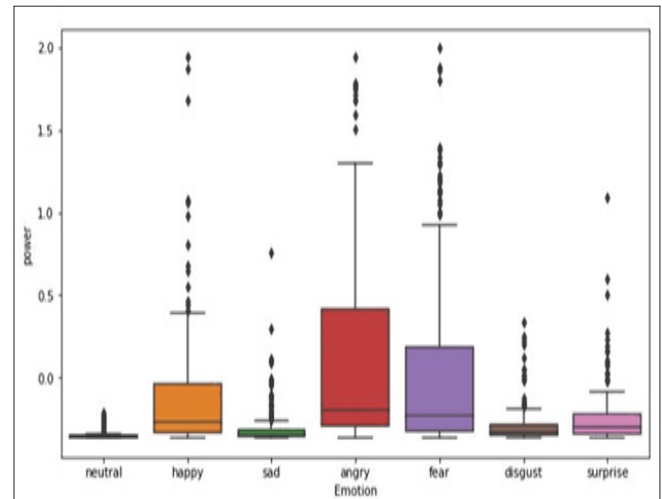


Figure 10. Distributions of Emotion with Respect to Gender<sup>12</sup>

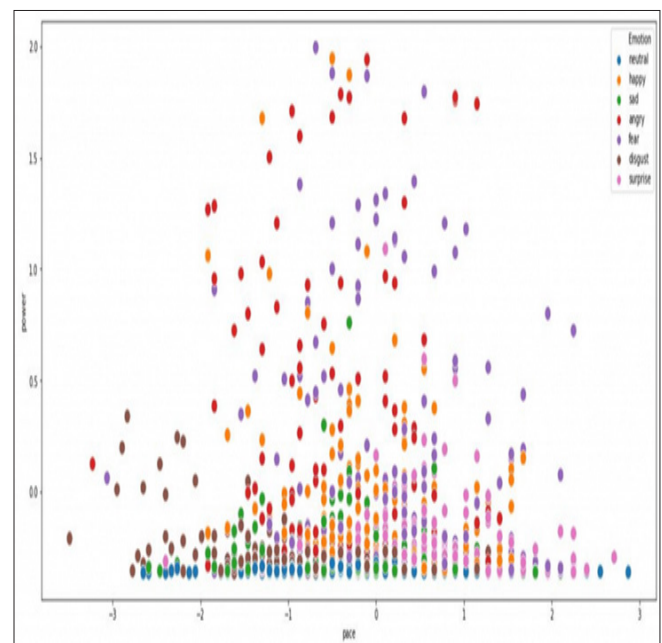


Figure 11. The Scatter of Power Vs Relative Pace of Audio Clips<sup>12</sup>

## Experimental Results

This section discusses the experiment outcomes. We demonstrate MLR, SVM, and RNN classifier accuracy. The experimental evaluation uses Spanish and Berlin datasets. Tenfold cross-validation yields all categorization results.

Performance analysis uses cross-validation. This method randomly divides data into N complementary subgroups. Each validation uses N1 of these subsets for training and the rest for testing. This approach uses a basic LSTM neural network. Two dense classification layers precede two successive LSTM layers with hyperbolic tangent activation. When applied to,<sup>6</sup> the SN method produced the best results. Each speaker's characteristics have a mean of 0 and a standard deviation of 1. Tables 2, 3, and 4 show the recognition rates from the Berlin and Spanish databases for the various attributes and classifiers. These investigations employ a feature set without feature selection. Table 2 shows that the SVM classifier performs better than 81% when using MFCC and MS for the Berlin database. We strengthened the model by modifying SVM parameters for each feature category, making our results better than.<sup>21</sup>

Recognition results with MS, MFCC features, and their combination on Berlin database; AVG. denotes average recognition rate;  $\sigma$  denotes standard deviation of the 10-cross-validation accuracies. Berlin(a, fear; e, disgust; f, happiness; l, boredom; n, neutral; t, sadness; w, anger).

Recognition results with MS, MFCC features, and their combination on Spanish database. Spanish (a, anger; d, disgust; f, fear; j, joy; n, neutral; s, surprise; t, sadness).

Recognition results using RNN classifier based on Berlin and Spanish databases.

Table 2, shows SN improves Berlin database recognition.

Tables 3 and 4, show that the Spanish database does not. Three classifiers provide identical results. Speaker counts explain this. Comparing the Spanish database to the Berlin database, which contains 10 speakers, shows linguistic effect. Spanish has two speakers. Table 4 shows that the Berlin database's RNN method's recognition rate is lowest when incorporating both types of attributes. The RNN model's 155 coefficients and poor training data cause this. Overfitting occurs. Table 5 shows an increase of almost 13% when we reduced the number of criteria from 155 to 59. We conducted all tests on the development set using recursive feature elimination (LR-RFE) for each modality combination to evaluate whether a smaller feature space improves recognition performance. Each iteration's feature ranking methodology affects RFE stability. Our sample RFE was examined using regression models and an SVM. Linear regression produced more consistent findings. Combining qualities yields the best results, as shown before. This combo employed just LR-RFE feature selection for accuracy. This research used 155 features and selected the best. LR-based RFE feature selection selected 59 Berlin database features and 110 Spanish database features. Table 5 compares LR-RFE results. LR-RFE does not improve accuracy in most scenarios using the Spanish database. Figure 12 shows that LR-RFE improves Berlin database recognition using the three classifiers. The RNN classifier's average of MFCC and MS features climbs from 63.67 to 78.11%. Table 5 shows these results. The Spanish database's best recognition rate is above 94.01% for MFCC and MS after RNN-LR-RFE selection.

Table 2.<sup>13</sup>

Test	Feature	Method	SN	Recognition Rate (%)							AVG.	(a)
				A	E	F	L	N	T	W		
#1	MS	MLR	NO	45.90	45.72	48.78	77.08	59.43	79.91	75.94	66.23	(5.85)
	MFCC			56.55	62.28	45.60	54.97	57.35	74.36	91.37	64.70	(3.20)
	MFCC+SM			70.26	73.04	51.95	82.44	69.55	82.49	76.55	73.00	(3.23)
#2	MS	SVM	NO	56.61	54.78	51.17	70.98	67.32	67.50	73.13	70.63	(6.45)
	MFCC			73.99	64.14	64.76	55.30	62.28	84.13	83.13	71.70	(4.24)
	MFCC+SM			82.03	68.70	69.09	79.16	76.99	80.89	80.63	81.10	(2.73)
#3	MS	MLR	YES	48.98	35.54	32.66	80.35	55.54	88.79	85.77	64.20	(5.27)
	MFCC			59.71	59.72	48.65	67.10	67.98	91.73	57.51	71.00	(4.19)
	MFCC+SM			72.32	68.82	51.98	82.60	81.72	91.96	80.71	75.25	(2.49)
	MS	SVM	YES	62.42	49.44	37.29	76.14	71.30	88.44	80.15	71.90	(2.38)
	MFCC			70.68	56.55	56.99	59.88	68.14	91.88	85.44	77.60	(4.35)
	MFCC+SM			77.37	69.67	58.16	79.87	88.57	98.75	86.64	81.00	(2.45)

Table 3.<sup>13</sup>

Test	Feature	Method	SN	Recognition Rate (%)							AVG.	(a)
				A	E	F	L	N	T	W		
#1	MS	MLR	NO	67.72	44.04	68.78	46.95	89.58	63.10	78.49	69.22	(1.37)
	MFCC			67.85	61.41	75.97	60.17	95.97	71.89	84.94	77.21	(0.76)
	MFCC+SM			78.75	78.18	80.68	63.84	96.80	82.44	89.01	83.55	(0.55)
#2	MS	SVM	NO	70.33	69.38	78.09	60.97	89.25	69.38	85.95	80.98	(1.09)
	MFCC			79.93	79.02	81.81	75.71	93.77	80.15	92.01	90.94	(0.93)
	MFCC+SM			84.90	88.26	89.44	80.90	96.58	83.89	95.63	86.89	(0.62)
#3	MS	MLR	YES	64.76	49.02	66.87	44.52	87.50	58.26	78.70	67.84	(1.27)
	MFCC			66.54	57.83	74.56	56.98	94.02	72.32	89.63	76.47	(1.51)
	MFCC+SM			77.01	78.45	80.50	64.18	94.42	80.14	91.29	83.03	(0.97)
	MS	SVM	YES	69.81	70.35	75.44	52.60	86.77	66.94	82.57	78.40	(0.95)
	MFCC			77.45	77.41	80.99	69.47	91.89	75.17	93.50	87.47	(0.95)
	MFCC+SM			85.28	84.54	84.49	73.47	93.43	81.79	94.04	86.57	(0.72)

Table 4.<sup>13</sup>

Dataset	Feature	SN	Average (avg)	Standard Deviation (a)
Berlin	MS	NO	66.32	5.93
	MFCC		69.55	3.91
	MFCC+MS		63.67	7.74
	MS	YES	68.94	5.65
	MFCC		73.08	5.17
	MFCC+MS		76.98	4.79
Spanish	MS	NO	82.30	2.88
	MFCC		86.56	2.80
	MFCC+MS		90.05	1.64
	MS	YES	82.14	1.67
	MFCC		86.21	1.22
	MFCC+MS		87.02	0.36

Table 5

SN	Classifier	LR-FREE	Berlin	Spanish
NO	MLR	NO	73.00 (3.23)	83.55 (0.55)
		YES	79.40 (3.09)	84.19 (0.96)
	SVM	NO	81.10 (2.73)	86.69 (0.62)
		YES	80.90 (3.17)	90.05 (0.80)
	RNN	NO	63.67 (7.74)	90.05 (1.64)
		YES	78.11 (3.53)	94.01 (0.76)
YES	MLR	NO	75.25 (2.49)	83.03 (0.97)
		YES	83.20 (3.25)	82.27 (1.12)
	SVM	NO	81.00 (2.45)	86.57 (0.72)
		YES	83.90 (2.46)	86.47(1.34)
	RNN	NO	76.98 (4.79)	87.02 (0.36)
		YES	83.42 (0.70)	85.00 (0.93)

Recognition results with a combination of MFCC and MS features using ML paradigm before and after applying LR-RFE feature selection method (Berlin and Spanish databases).<sup>13</sup>

## Conclusion

We created an autonomous Speech-Emotion Recognition (SER) system for this study. This system categorizes seven

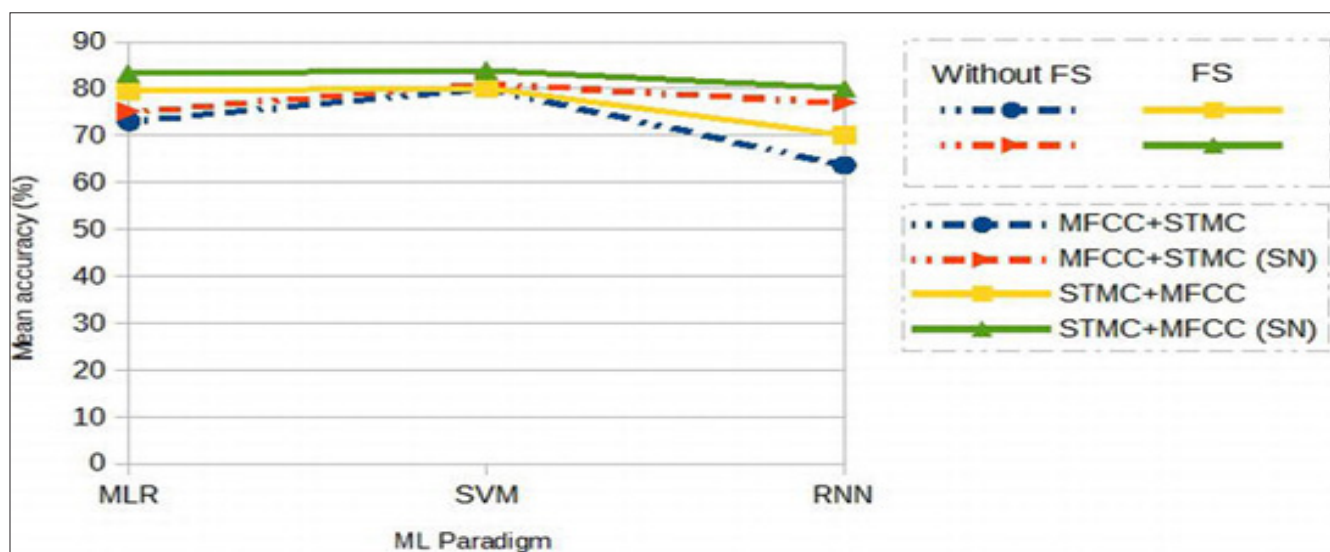


Figure 12.<sup>13</sup>

Table 6.<sup>13</sup>

Emotion	Anger	Disgust	Fear	Joy	Neutral	Surprise	Sadness	Rate %
Anger	79	1	0	1	2	3	0	91.86
Disgust	0	67	3	0	1	0	1	93.05
Fear	0	3	70	0	1	0	2	93.33
Joy	3	1	1	71	0	0	0	93.42
Neutral	2	0	1	0	156	0	1	97.50
Surprise	2	1	0	3	0	60	0	92.30
Sadness	0	0	1	0	2	0	66	95.65
Precision %	91.86	91.78	92.10	94.66	96.29	95.23	94.28	

Performance comparison of three Machine Learning Paradigms (MLR, SVM, RNN) using Speaker Normalization (SN) and RFE Feature Selection (FS), for the Berlin database, is shown.

The confusion matrix for the best recognition of emotions using MFCC and MS features with RNN based on the Spanish database is shown in 6. The rate column lists per class recognition rates and precision for a class are the number of samples correctly classified divided by the total number of samples classified to the class. It can be seen that Neutral was the emotion that was least difficult to recognize from speech as opposed to Disgust which was the most difficult and it forms the most notable confusion pair with Fear.

Confusion matrix for feature combination after LR-RFE selection based on Spanish database.

sensations using MLR, SVM, and RNN. MFCC and MS characteristics from the Berlin and Spanish databases were used to combine these qualities. We examine how classifiers and features may enhance spoken language emotion recognition. Selecting a subset of traits that distinguish people well. Feature selection methods in machine learning show that more information is not necessarily better. Using these traits, machine learning models were created and tested to categories emotional states. SER claimed the best Spanish database recognition rate of 94% utilising an RNN classifier without speech normalization (SN) and Feature Selection (FS). After speaker normalization (SN) and Feature Selection (FS), all Berlin database classifiers reach 83% accuracy. This shows that RNNs perform better with more data but have very lengthy training cycles. As a result, the SVM and MLR models had a larger potential

for real-world applicability with less data than the RNN model. Combining classifiers and databases may increase emotion identification system resilience. A single detection system can examine the impact of training several emotion detectors. Since a good emotion feature selection technique can rapidly locate features expressing emotional states, we also try to apply additional feature selection methodologies. We employ these tactics. Our goal is to create a classroom pedagogical system. This helps teachers organise their courses. We'll test this research system to achieve our goal. Finally, the delta features from each phrase are extracted before categorising the data using the SVM hierarchical framework. Results may be statistically significant even if the standard deviation is more than one.

## References

1. Alexander I. Iliev, Michael S. Scordilis, Joao P, et al. Falcao, "Spoken emotion recognition through optimum-path forest classification using glottal features", *Computer Speech and Language*. 2010;24:445-460.
2. Ashish B. Ingale and Dr.D.S.Chaudhari, "Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine", *International Journal of Advanced Engineering Research and Studies*. 2012;1.
3. Enrique M. Albornoz, Diego H. Milone and Hugo L. Rufiner, "Spoken emotion recognition using hierarchical classifiers", *Computer Speech and Language*. 2011;25:556-570.
4. Enrique M. Albornoz, Diego H. Milone and Hugo L. Rufiner, "Spoken emotion recognition using hierarchical classifiers", *Computer Speech and Language*. 2011;25:556-570.
5. Kwon, Oh-Wook, Chan, et al. "Emotion recognition by speech signals", *Eurospeech - Geneva*. 2003;125- 128.
6. T. Bänziger, K. R. Scherer, "The role of intonation in emotional expression", *Proc. IEEE Int'l Conf. on Speech Communication*. 2005;46:252-267.
7. C. Busso, S. Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", *IEEE Trans. on Audio, Speech and Language processing*. 2009;17(4):582-596.
8. Kyung Hak Hyun, Eun Ho Kim, Yoon Keun Kwak, "Improvement of Emotion Recognition by Bayesian Classifier Using Nonzero-pitch Concept". 2005;7803-9275-2/05/2005 IEEE.
9. Eun Ho Kim, Kyung Hak Hyun, "Robust Emotion Recognition Feature, Frequency Range of Meaningful Signal" *IEEE International Workshop on Robots and Human Interactive Communication*, 0-7803-9275-2/05 2005IEEE (2005).
10. G. Zhou, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech, Audio Proc.* 2001;9:201-216.
11. TS Polzin A. Waibel, "Emotion-sensitive humancomputer interfaces," *ISCA Workshop, Speech and Emotion*, 2000.
12. <https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/>
13. <https://www.intechopen.com/chapters/65993>
14. Peipei S, Zhou C, Xiong C. Automatic speech emotion recognition using support vector machine. *IEEE*. 2011;2:621-625.
15. Sathit P. Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. *International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2015:73-76.
16. Alex G, Navdeep J. Towards end-to-end speech recognition with recurrent neural networks. In: *International Conference on Machine Learning*. 2014;32.
17. Chen S, Jin Q. Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Australia: Brisbane; 2015.
18. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. *Asia-Pacific*. 2017:1-4.
19. Sara M, Saeed S, Rabiee A. Speech Emotion Recognition Based on a Modified Brain Emotional Learning Model. *Biologically inspired cognitive architectures*. Elsevier; 2017;19:32-38.
20. Yu G, Eric P, Hai-Xiang L, van den HJ. Speech emotion recognition using voiced segment selection algorithm. *ECAI*. 2016;285:1682-1683.