ISSN No. 0976-5697

VOLUME 12 SPECIAL ISSUE 2, JUNE 2021

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AUTOMATIC SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

Meshach A Martin
Department of Computer Science & Engineering
REVA University
Bangalore, India
R18CS226@cit.reva.edu.in

G Sai Charith
Department of Computer Science & Engineering
REVA University
Bangalore, India
R18CS151@cit.reva.edu.in

Salony Shah
Department of Computer Science & Engineering
REVA University
Bangalore, India
R18CS2529@cit.reva.edu.in

Nehal Singh
Department of Computer Science & Engineering
REVA University
Bangalore, India
R14CS123@cit.reva.edu.in

Prof. Dasari Bhulakshmi
Bhulakshmi.d@reva.edu.in
Department of Computer Science & Engineering
REVA University
Bangalore, India

Abstract: For several years, emotion detection from speech signals has been a research topic in human-machine interface applications. To discern emotions from speech signals, a variety of devices have been developed. Theoretical definitions, categorizations, and modalities of emotion expression are all discussed.

To conduct this research, a SER framework based on various classifiers and feature extraction methods was developed. The mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) characteristics of speech signals are analysed and fed into various classifiers for training. Using feature selection, this method is used to find the most important function subset (FS). The features extracted from emotional speech samples that make up the database for the speech emotion recognition system include power, pitch, linear prediction cepstrum coefficient (LPCC), and Mel frequency cepstrum coefficient (MFCC). The effectiveness of classification is determined by the extracted features.

Seven emotions are classified using a recurrent neural network (RNN) classifier. Their results are then compared to techniques such as multivariate linear regression (MLR) and support vector machines that are used in the field of emotion detection for spoken audio signals (SVM).

Keywords: Modulation spectral (MS), Mel-frequency cepstrum coefficients (MFCC), Feature selection (FS), Linear prediction cepstrum coefficient (LPCC), Recurrent Neural Network (RNN) Classifier, Multivariate linear regression (MLR).

I. INTRODUCTION

There are many ways to communicate, but one of the fastest and most natural is through voice. As a consequence, speech can be a fast and efficient means of communication between humans and machines. Humans have the inherent ability to fully comprehend the message they are receiving by using all of their senses. There has been an increase of research interest in this field over the last few years. Emotions can be observed in a number of situations, such as robot interfaces, audio monitoring, web-based E-learning, business applications, clinical trials, entertainment, banking, call centres, cardboard structures, video games, and so on.

The emotions conveyed by male and female speakers are discovered in speech emotion recognition. Fundamental frequencies, Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC), and other speech features were studied in the previous century, and they are still used to process speech today. On the other hand, using a combining feature set will result in high dimension and redundancy of speech features, complicating the learning process for most machine learning algorithms and raising the risk of overfitting. As a consequence, feature selection is needed to reduce redundancy in feature dimensions.

Recognizing emotions from a speaker's speech is highly difficult due to the following factors: It's uncertain which aspects of speech are more effective at distinguishing between different emotions.

The speech emotion recognition system can be used for psychological diagnosis, intelligent toys, lie detection, and call centre conversations, among other things.

II.BACKGROUND

Emotion recognition and affective computing on vocal social media:

While vocal messages are typically used to express semantic information, they also contain a significant amount of emotional data, which is a new subject for social media analytics data mining. Using a computational approach for emotion detection and affective computing on vocal social media, this paper proposes a method for estimating complex emotion and its dynamic shifts in a three-dimensional PAD (Panel of Affective Display).

Speaker-sensitive emotion recognition via ranking:

We offer an emotion detection ranking scale that includes facts about the speakers' total expressivity naturally. We show that, as compared to traditional methods, our methodology significantly improves accuracy. By treating each speaker's data as a separate question and combining the predictions of all rankers, we train ranking SVMs for individual emotions. The ranking system seems to have two advantages. Even in speaker-independent training/testing settings, it captures speaker-specific information.

Using a hierarchical binary decision tree approach to recognise emotions:

In order to increase analytical capacity and build human–machine interfaces that make for efficient communication, it's important to develop reliable and dependable emotion detection systems that are suitable for real-world applications. To consider emotions, we use a hierarchical computational model. Using successive levels of binary classifications, the proposed system maps an input speech utterance into one of the multiple emotion groups. The main concept is that the stages of the tree are organised in such a way that the most simple classification tasks are tackled first, preventing error propagation.

III.PROPOSED TOPIC

Speech Emotion Recognition System:

The pattern recognition task is very similar to the strategy for recognising speech emotions. Understanding the stages involved in speech emotion recognition is aided by the pattern recognition cycle, which is dependent on the modular flow of signal data in various stages. The features are created by extracting pitch MFCCs and Wavelet domain knowledge. Following that is the feature selection process, which limits the coefficients of the feature set to reduce the curse of dimensionality while maintaining the associated features. The selected features are then fed into Gaussian Mixture Model (GMM) and K-Nearest Neighbour classifiers (K-NN).

IV. DESIGN AND IMPLIMENTATION

Block diagram

There are four key phases in our SER scheme. The first is a series of voice samples. The second features vector is generated after the features have been extracted. As a next step, we attempted to decide which characteristics are most important in distinguishing each emotion. For identification, these features are applied to a machine learning classifier.

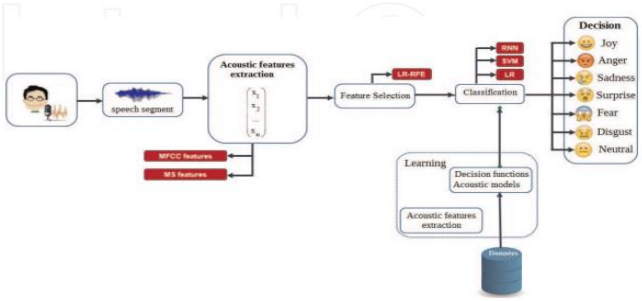

Figure 1.
Block diagram of the proposed system.

Feature extraction

The emotional characteristics are described by a large number of parameters in the speech signal. Choosing which features to use is one of the most difficult aspects of emotion detection. Recent research has extracted many common features such as energy, pitch, and formant, as well as some spectrum features such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features. In this analysis, we used modulation spectral features and MFCC to extract emotional features.

The most widely used tool for describing the spectral properties of speech signals is the Mel-frequency cepstrum coefficient (MFCC). They are the best for speech recognition because they take into account human interpretation of frequencies. The Fourier transform and energy spectrum for each frame is computed and projected onto the Mel-frequency scale. The Mel log energies were discrete cosine transformed (DCT), and the first 12 DCT coefficients were used to calculate the MFCC values for the classification. The formula for calculating MFCC is seen in the diagram.

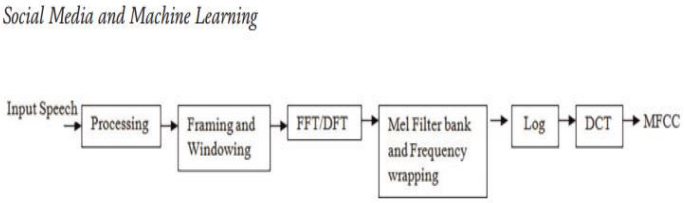

Figure 2.
Schema of MFCC extraction [33].

In our study, we extract the first 12 orders of the MFCC coefficients from speech signals sampled at 16 KHz. For each order coefficient, we calculate the mean, variance, standard deviation, kurtosis, and skewness, and this applies to all frames of an utterance. The MFCC has 60 dimensions for each function vector.

Modulation spectrum features MSFs are constructed from a long-term spectro-temporal representation inspired by auditory feedback. These properties are obtained by simulating spectro-temporal (ST) processing in the human auditory system while taking into account natural acoustic and modulation frequencies. The measures for measuring the ST representation are depicted in the diagram below. To achieve the ST representation, an auditory filterbank first decomposes the speech signal (19 filters in total). Modulation signals are produced by computing the Hilbert envelopes of the critical-band outputs. A modulation filterbank is then used to analyse the Hilbert envelopes for frequency. Modulation spectra are the spectral contents of modulation signals, and as a result, the proposed features are known as modulation spectral features (MSFs). Finally, the ST representation is calculated by calculating the energy of the decomposed envelope signals as a function of standard acoustic and modulation frequencies. The energy, which is calculated over all frames in each spectral band, provides a function. An auditory filterbank with N ¼ 19 filters and a modulation filterbank with M ¼ 5 filters were used in our experiment. In sum, 95 (19X5) MSFs are calculated from the ST representation in this paper.

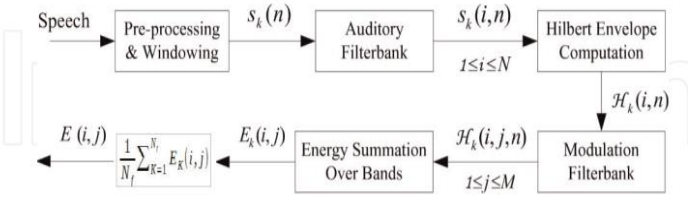

Figure 3.
Process for computing the ST representation [5].

Feature selection

The aim of feature selection in machine learning is to "reduce the number of features used to characterise a dataset in order to increase the efficiency of a learning algorithm on a specific task." The aim will be to improve classification accuracy in a particular task using a specific learning algorithm while reducing the amount of features used to generate the final classification model. Feature selection (FS) is a method for choosing a subset of relevant features from a wider range based on a significance assessment criterion, which typically leads to improved identification accuracy. It has the ability to reduce the time it takes for learning algorithms to run dramatically. In this part, we explain the LR-RFE that we used in our study.

Classification methods

A variety of machine learning algorithms have been used to detect distinct emotions. This algorithms are made to learn from training samples before using what they've observed to find new findings. In reality, there is no one-size-fits-all solution to the issue of which learning algorithm to use; each technique has its own set of advantages and disadvantages. As a consequence, in this analysis, we agreed to analyse the quality of three different classifiers.

Multivariate linear regression classification (MLR) is a computing approach for machine learning algorithms that can be used for both regression and classification problems. Algorithm 1's LRC algorithm has been modified somewhat. The absolute value of the difference between the original and projected response vectors was determined as ($|y - y_i|$), instead of the Euclidean distance between them ($\|y - y_i\|$).

In machine learning, support vector machines (SVM) are an optimal margin classifier. It's also been used in a host of experiments on audio emotion recognition.

Algorithm 1. Linear Regression Classification (LRC)

Inputs: Class models $X_i \in \mathbb{R}^{q \times p_i}$, $i = 1, 2, ..., N$ and a test speech vector $y \in \mathbb{R}^{q \times 1}$
Output: Class of y

1. $\hat{\beta}_i \in \mathbb{R}^{p_i \times 1}$ is evaluated against each class model, $\hat{\beta}_i = (X_i^T X_i)^{(-1)} X_i^T y$, $i = 1, 2, ..., N$
2. $\hat{y}_i$ is computed for each $\hat{\beta}_i, \hat{y}_i = X_i \hat{\beta}_i, \quad i = 1, 2, ..., N$;
3. Distance calculation between original and predicted response variables $d_i(y) = |y - y_i|, \quad i = 1, 2, ..., N$;
4. Decision is made in favor of the class with the minimum distance $d_i(y)$

Recurrent neural networks (RNN) have shown to increase classification accuracy and are well suited to learning time series results.

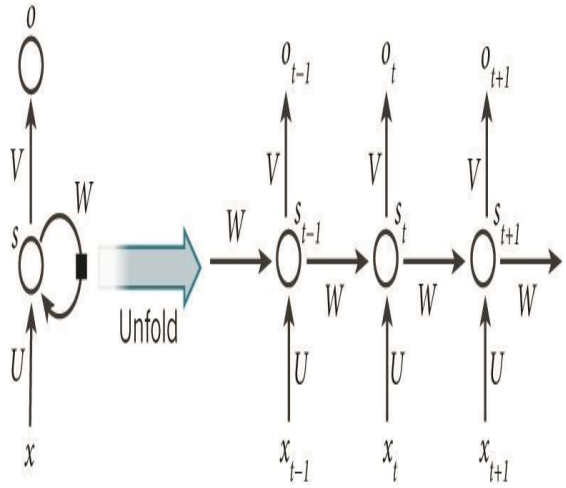

Figure 4.
A basic concept of RNN and unfolding in time of the computation involved in its forward computation [18].

A basic RNN implementation definition is seen in Figure 4. Unlike traditional neural networks, which use different parameters for each layer, the RNN uses the same parameters in all phases (U, V, and W). The following are the hidden state formulas and variables:

$$s_t = f(Ux_t + Ws_{t-1})$$

Where x_t, s_t, and o_t are respectively the input, the hidden state, and the output at time step t and U,V,W are parameters matrices.

Emotional speech databases:

The reliability and robustness of recognition systems can be quickly harmed if they are not well trained with an appropriate database. As a result, including adequate and sufficient phrases in the corpus to train and validate the emotion detection system is crucial. The three main types of databases are acted emotions, natural random emotions, and elicited emotions. In this analysis, we used acted emotion databases because they contain a large number of strong emotional expressions. According to the literature, the bulk of studies on speech emotion detection has used emotional acted speech. The Berlin Database and the Spanish Database are the two emotional expression databases that we used in our studies to describe different emotions.

Berlin database

The Berlin database is widely used in emotional expression recognition. It includes 535 lines of dialogue delivered by ten actors (5 females, 5 males) in seven distinct emotional states (anger, boredom, disgust, fear, joy, sadness, and neutral). This database was chosen for two reasons: I it has excellent recording quality, and (ii) it is a commonly used and public database for emotion detection that has been recommended in the literature.

Spanish database

Two seasoned actors' utterances are included in the INTER1SP Spanish emotional archive (one female and one male speaker). The (six basic emotions plus neutral (anger, sorrow, joy, terror, disgust, surprise, and neutral/normal)) were recorded twice in the Spanish corpus that we have access to (free for academic and study use). Four new neutral variants have been discovered (soft, noisy, slow, and fast). This database was chosen over others because it is more user-friendly and provides researchers with more information (6041 utterances in total). To achieve a higher and more precise rate of identification and to enable comparison with the Berlin database mentioned above, this paper concentrated on only seven key emotions from the Spanish database.

V.RESULTS

The results of the experiments are summarised and discussed in this section. The identification accuracy of the MLR, SVM, and RNN classifiers is evaluated. On the Berlin and Spanish databases, an experimental assessment is carried out. Tenfold cross-validation is used to collect all classification results. In cross-validation, data is randomly divided into N complementary subsets, one for preparation and the other for checking in each validation. A basic LSTM neural network architecture was used. It consists of two thick grouping layers followed by two LSTM layers with hyperbolic tangents. Data attributes are scaled to 12 1; 1 before classifiers are implemented.

In comparison to previous results, we strengthened our results by changing the SVM parameters for each form of operation to create a good model. As seen in Table 1, using SN boosts identification results for the Berlin index. Tables 2 and 3 show that this is not the case with the Spanish database. The three respective classifiers produce similar outcomes. The number of speakers in each database would demonstrate this. The Berlin database has ten distinct speakers, while the Spanish database only has two, indicating that the language influence is more likely. For the Berlin database, we found that combining all types of functionality in the RNN form has the lowest identification rate. (See Table 3 for details). This is due to the RNN model's excessive number of parameters and low performance. Overfitting is the term for this occurrence. This is supported by the fact that when the number of features was decreased from 155 to 59, the results increased.

				Recognition rate (%)								
Test	Feature	Method	SN	A	E	F	L	N	T	W	AVG.	(σ)
#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	87.51	71.00	(4.19)
	MFCC+SM			72.32	68.82	51.98	82.60	81.72	91.96	80.71	75.25	(2.49)
#4	MS	SVM	Yes	62.72	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)

Berlin (a, fear; e, disgust; f, happiness; l, boredom; n, neutral; t, sadness; w, anger).

Table 1: Recognition results with MS, MFCC attributes, and their combination on the Berlin database; AVG. denotes the

average identification rate; SD denotes the standard deviation of the 10-cross-validation accuracies.

				Recognition rate (%)								
Test	Feature	Method	SN	A	D	F	J	N	S	T	AVG.	(σ)
#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	95.79	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	89.69	(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)
#4	MS	SVM	Yes	69.81	70.35	75.44	52.60	86.77	66.94	82.57	78.40	(1.64)
	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)

Spanish (a, anger; d, disgust; f, fear; j, joy; n, neutral; s, surprise; t, sadness).

Table 2: shows the impact of recognition on the Spanish database using MS, MFCC, and a combination of these functions.

RFE's stability is influenced by the type of function rating model used at each iteration. We used an SVM and regression models to evaluate the RFE in our case; we found that linear regression yields more stable outcomes. We can see from the previous findings that combining the features yields the best results. We only used LR-RFE feature selection for this combination to boost performance. This research used a total of 155 features.

Dataset	Feature	SN	Average (avg)	Standard deviation (σ)
Berlin	MS	No	66.32	5.93
	MFCC		69.55	3.91
	MFCC+MS	Yes	63.67	7.74
	MS		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 3: RNN classifier recognition results based on Berlin and Spanish databases.

VI.FUTURE WORK

Increased collaboration compatibilities with applications having a virtual teaching experience. Increased collaboration compatibilities with applications of therapeutic analysis and aid. A machine to understand and set ambient environment required by the user. A collaboration with an IOT based machine running on an arduino board to set these physical environmental conditions.

VII.CONCLUSION

The use of multiple classifiers to recognise speech emotion is demonstrated. Two critical issues in speech emotion recognition systems are the signal processing unit, which derives appropriate functionality from usable speech signals, and the classifier, which recognises emotions from the speech signal. The accuracy of most classifiers in a speaker independent system is lower than in a speaker based system.

The robustness of an emotion processing system can also be improved by combining databases and classifiers. The effect of training multiple emotion detectors can be investigated by integrating multiple emotion detectors into a single detection system. We may want to use other feature selection approaches because the quality of feature selection has an impact on the rate of emotion recognition: a successful emotion feature selection system can quickly pick features that fit the emotion state.

VIII.REFERENCES

[1] Ali H, Hariharan M, Yaacob S, Adom AH. Facial emotion recognition using empirical mode decomposition. Expert Systems with Applications. 2015;42(3): 1261-1277

[2] Liu ZT, Wu M, Cao WH, Mao JW, Xu JP, Tan GZ. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing. 2018;273: 271-280

[3] Ragot M, Martin N, Em S, Pallamin N, Diverrez JM. Emotion recognition using physiological signals: Laboratory vs. wearable sensors. In: International Conference on Applied Human Factors and Ergonomics. Springer; 2017. pp. 15-22

[4] Surabhi V, Saurabh M. Speech emotion recognition: A review. International Research Journal of Engineering and Technology (IRJET). 2016;03:313-316

[5] Wu S, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication. 2011;53:768-785

[6] Wu S. Recognition of human emotion in speech using modulation spectral features and support vector machines [PhD thesis]. 2009

[7] Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data Classification: Algorithms and Applications. 2014:37

[8] Martin V, Robert V. Recognition of emotions in German speech using Gaussian mixture models. LNAI. 2009; 5398:256-263

[9] Ingale AB, Chaudhari D. Speech emotion recognition using hidden Markov model and support vector machine. International Journal of Advanced Engineering Research and Studies. 2012:316-318

[10] Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

[11] Milton A, Sharmy Roy S, Tamil Selvi S. SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications. 2013;69.

