Implementation of SVM with SMO for Identifying Speech Emotions using FFT and Source Features

Sheetal Patil
Research Scholar
Department of Electronics and Telecommunication Engineering
Matoshri College of Engineering & Research Center
Nashik, India
sheetalsp2105@gmail.com

Dr. G. K. Kharate
Professor & Principal
Department of Electronics and Telecommunication Engineering
Matoshri College of Engineering & Research Center
Nashik, India
gkkharate@rediffmail.com

Abstract: For speaker independent emotion recognition, scholars have recently proposed Fourier parameter and mean Fourier parameter models. This study aims to propose fast Fourier coefficient features such as minimum, maximum, mean, and standard deviation. In addition to source features, i.e., glottal volume velocity, this study aims to separate speech segments into voiced speech segments, unvoiced speech segments, and silent segments so that the effect of each part of speech corpus on emotion recognition can be observed. Experimental results indicate that the proposed method improves speech emotion recognition rate to 80.85% for EmoDB. For Ryerson audio-visual database of emotional speech and song (RAVDESS) for eight emotions, a highest emotion recognition rate of 70.19% was achieved.

Keywords: emotions, fast Fourier transform features, recognition, support vector machine, SMO

1. Introduction

Emotion recognition technologies have evolved from a specialty to a major component of human–computer interaction (HCI). Such technologies aim to make natural contact with computers possible by using direct speech interaction instead of conventional interfaces that need an input to interpret verbal contents such that a human audience can respond more quickly. Dialogue technologies for spoken languages, such as a call center chat, an onboard car driving device, and the use of emotion patterns from voice in the medical sector are examples of implementations that use such systems. Nevertheless, there are several problems that exists in HCI systems that need to be addressed appropriately, especially when such systems have a transition from a laboratory testing to real-world applications, and so, a lot of efforts are needed to efficiently handle such problems and achieve superior emotion recognition by machines. Speech emotion recognition (SER) is used in various applications such as text to speech synthesis, education, automobile industries, medical diagnosis, and multimedia content management.

Speech is a multi-dimensional signal that includes details about the message, the speaker, gender, vocabulary, and emotions. It's made by a vocal tract (VT) device that's excited by a time-varying excitation stream. Excitation signals produce vocalization, whispering, frication, compression, and vibration. On the basis of excitation, a speech signal can be categorized as voiced and unvoiced segments. For a voiced signal, excitation is periodic that is generated via the fundamental frequency of vocal cords. For an unvoiced signal, fundamental frequency does not exists in an excitation signal; so, there is a need to consider white noise as excitation [1].

The challenge of determining a human's emotional state is unique and could serve as a benchmark for any emotion detection model. A differentiated emotional approach is considered one of the central approaches among the various models used for the categorization of emotions. Anger, boredom, disgust, surprise, terror, excitement, satisfaction, neutrality, and sadness are among the emotions included. A three-dimensional continuous space with parameters such as arousal valence and power is another important model that is used.

The approach for SER predominantly comprises two phases, i.e., feature extraction phase and feature classification phase. In the first phase—which is the speech processing field—researchers have drawn several features such as source features, prosodic features, and system features, whereas the second phase consists of feature classification via linear and nonlinear classifiers. Note that Bayesian networks or the maximum likelihood principle, hidden Markov model (HMM), support vector machines (SVMs), random forest, convolutional neural networks (CNNs), and Gaussian mixture models (GMMs) are some of the widely used classifiers for emotion recognition. As speech signals are nonstationary, nonlinear classifiers function efficiently for SER.

The remainder of this paper is organized as follows. Section II focuses on the related work that is performed in this field. Section III explicates on the two databases that are used in this study. Section IV describes about the

2. Related Work

Krothapalli and Koolagudi elaborated on excitation source features, LPPCC residual, LP residual signal phase, instants of glottal closure and glottal volume velocity (GVV) parameters. With the help of auto-associative neural networks (AANNs) and SVMs, they achieved an average accuracy of 62.33% with LP residual samples around GCI for the Berlin database and 63.33% for the Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) database [1].

Koduru and others focused on RAVDESS in English. Features extracted were prosodic features, i.e., pitch, energy, and ZCR and system features, i.e., DWT and MFCC. The classifiers used were LDA, decision tree, and SVM. They achieved 85% accuracy by means of decision tree and 70% accuracy via SVM but only for four classes of emotion [2].

Venkataramanan and Rajmohan performed SER experiments with RAVDESS by considering features like Log-Mel Spectrogram, MFCC, pitch, and energy. The implication of features for emotion classification was compared by applying methods such as CNNs, HMM, and deep neural networks (DNNs). For 14 classes (i.e., two genders and seven emotions), they achieved an accuracy of 68% with CNN [3]

Koolagudi, Murthy, and Bhaskar discussed about three classifiers, namely k-means clustering, vector quantization, and ANNs. Prosodic features such as pitch, intensity, jitter, and shimmer, spectral features such as MFCC, and formants were considered. The authors combined statistical parameters with spectral features to acquire superior performance. Moreover, they achieved an overall accuracy of 81% on the IITKGP-SESC database with five classes of emotion [4].

Iqbal and Barua used two datasets, i.e., SAVEE and RAVDESS, to train the SER system. A total of 34 audio features such as MFCC, energy, spectral entropy, etc. was extracted. Classifiers such as k-NN, GB, and SVM were used for four emotions [5].

Hao, Tianhao, and Fei were of the opinion that an optimization algorithm with a good classifier would largely improve the rate of recognition, and so, they proposed a nonlinear SVM algorithm with SMO optimization. For EmoDB, they achieved an accuracy of 85.15% for five emotions [6].

Ghai et al. [7] worked with EmoDB with MFCC and energy features. They used SVM, GB, and random forest classifiers and attained an accuracy of 55.89%, 65.23%, and 81.05%, respectively.

Wang et al. [8] proposed a new Fourier parameter model by means of perceptual content of voice quality and first- and second-order differences for a speaker-independent SER. As feature vectors, they looked at 120 harmonic coefficients, their first- and second-order variations, as well as their minimum, mean, median, and standard deviation. The used SVM as the classifier. Nonetheless, it was observed that the acoustic features that were extracted had many dimensions, which led to a heavy computation burden. A recognition rate of 79.51% was attained using EmoDB through six emotions.

Chen et al. [9] used mean Fourier parameter for 120 STFT of a frame, which indeed reduced an acoustic feature's dimension to a great extent. Through MFP, an accuracy of 48% was attained by means of a k-NN classifier, 81% by means of a decision tree, and 88% by means of a random forest for EmoDB.

Kerkeni and others [10] extracted MFCC and modulation spectral features, and with EmoDB and SVM, they acquired a recognition rate of 59.50%.

3. Databases

In the reviewed literature, it was observed that emotional speech databases were collected by means of the following methods: (a) actors were requested to portray the given emotion in the case of simulated databases, which were the most popular databases that were used in an emotional speech research, (b) elicited databases were not completely natural but were recorded under simulated natural situations, and (c) naturalistic databases were recorded from natural situations. According to the reviewed literature, there is a huge disparity between databases in terms of vocabulary, number of feelings, number of topics, and corpus selection intent, as well as database collection methods. It's worth noting that English led the emotional expression datasets, followed by German and Chinese. Furthermore, databases in languages such as Russian, Dutch, Slovenian, Swedish, Japanese, and Spanish were scarce. We choose RAVDESS in the English language and EmoDB in the German language for this review.

3.1 RAVDESS

It consists of 7356 files, which accounts for a total size of 24.8 GB. The database contains 24 trained actors (eighteen males and eighteen females) who vocalize two lexically matched phrases in a neutral North American accent. Speech comprises calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression was produced at two levels of emotional intensity (i.e., normal and strong) with an additional neutral expression. All conditions were made available in three modality formats, i.e., audio only (16 bit, 48 kHz .wav), audio video (720 p H.264, AAC 48 kHz, .mp4), and video only (no sound). Moreover, a further set of seventy-two participants offered test–retest data. High levels of emotional validity, interrater reliability, and test–retest intrareader reliability were reported. Validation data was open access and could be easily downloaded. From complete database for this research, only speech file (i.e., audio_speech_actors_01-24.zip, 215 MB) containing 60 trials per actor × 24 actors = 1440 files was used [11].

3.2 EmoDB

It is a freely available German emotional database that was created by the Institute of Communication Science, Technical University, Berlin, Germany. A total of ten professional speakers (five males and five females) participated in data recording. The database comprises 535 utterances and seven emotions, i.e., anger, boredom, anxiety, happiness, sadness, disgust, and neutral. Ten different sentences were used to record. The data was recorded at a sampling rate of 48 kHz and then down sampled to 16 kHz.

4. Classifier:

Apart from feature extraction, relevant classifiers are needed for emotion recognition. Emotions may be recognized using a variety of classifiers. Choosing the right classifier is often a trial-and-error process. The bulk of the time, a classifier's decision is based on previous advice [12]. Since the majority of function vectors are linearly inseparable in real-world contexts, a nonlinear classifier is a safer choice [13]. Nonparametric machine learning algorithms are those that do not make firm assumptions about the shape of a mapping element. These algorithms are free to learn any practical type from training data because they don't make assumptions. Nonparametric algorithms are useful because there is a lot of data and little prior experience, and we don't want to spend too much time figuring out which functionality to use. In building the mapping function, nonparametric algorithms aim to match the training data as closely as possible, retaining some capacity to generalize unseen data. These algorithms can be used in a number of ways. In this study, SVM is selected as the nonparametric algorithm.

4.1 SVM

Corinna Cortes and Vladimir Vapnik proposed it in the 1990s, and it is widely used in pattern recognition. SVM is used in regression and classification because of its good nonlinear fitting ability. SVM may thus be used to describe feelings. If a kernel obeys Mercer's theorem for nonlinear problems, it corresponds to the inner product of a particular transformation space. With the right kernels, nonlinear vectors can be converted to linear vectors with minimal computing overhead. A kernel is the dot product of a vector. A similarity or distance measure between new data and support vectors is defined by the kernel. Different kernels used in SVM are linear, polynomial, and radial basis function (RBF). Polynomial and RBF kernels transform an input space into higher dimensions. It is desirable to use more complex kernels as it allows lines to separate the classes that are curved or even more complex. In this study, RBF is used as the kernel. RBF is given as

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_\mathrm{i}) = \mathrm{e}^{-\mathrm{gamma}\ \sum(\mathrm{x}-\ \mathrm{x}_\mathrm{i}{}^2)} \qquad (1)$$

The gamma value chosen here is 0.1. There are two different families of solution aiming to extend SVM for multiclass problems [14]. The first solution follows the strategy of “one against one”, whereas the second solution follows the strategy of “one against all”. In this study, we use the strategy of “one against one”. Moreover, in this study, the nonlinear SVM method is used to classify a task based on a sequence minimization optimization algorithm that preserves the internal structure of data completely and improves classification accuracy. SMO algorithm is characterized by decomposing an original quadratic programming problem into a quadratic programming subproblem with only two variables and solving it until all variables satisfy the KKT condition [6]

5. Feature Extraction:

5.1Framing:

Speech is the most common way for humans to exchange information. It has a 4 kHz bandwidth and can communicate information with the emotion of a human voice. The following are some of the features of a voice signal. Since it is an independent variable, it is a one-dimensional signal with respect to time. It's unpredictable and non-stationary. Its frequency range also changes over time. Human voice has major frequency components just up to 4 kHz, despite having an audible frequency spectrum of 20 Hz–20 kHz. The method of partitioning a continuous speech signal into fixed length fragments is known as signal framing [15]. It is essential to make a speech signal stationary so that feature extraction from the stream of audio signal becomes simple and frame analysis can be performed independently. In this study, frames are created with 50% overlapping to avoid information loss. Herein, we have used a frame of 20ms with an overlapping of 10 ms. Therefore, each frame of RAVDESS comprises 960 samples/frame with an overlapping of 480 samples. EmoDB comprises 320 samples/frame with an overlapping of 160 samples. During framing, discontinuity takes place at the edges of each frame. This discontinuity causes several problems ahead during Fourier analysis. As a result, a tapered window is used to reduce spectral leakage. In this study, Hamming window is used, which is given by

$$w = [n]\ 0.54 - 0.46 \cos(\frac{2\pi n}{N-1}),\ 0 \leq n \leq N-1 \qquad (2)$$

5.2 Voiced, Unvoiced, and Silence Separation:

There are two sections of a basic speech phrase. One contains speech content, while the other contains silent or noise portions that occur between utterances but do not include any verbal information. There are two forms of articulated and unvoiced speech in the vocal portion of speech. Vowel sounds are used in spoken expressions. It is created by pressing air into the glottis, adjusting the friction of the vocal cords to open and close them, and producing all intermittent air pulses. The pulses excite the VT. Most of the information is carried out by the voiced segment. Since unvoiced expression lacks a fundamental frequency in an excitation signal, excitation is regarded as white noise. A VT constriction occurs at many points between the glottis and the lips, forcing air flow. Certain sounds are produced when a complete stoppage of air flow is followed by a sudden release. This generates an impulsive turbulent excitation that is repeatedly followed by additional turbulent excitation. The steadiness of unvoiced speech is less as compared to that of voiced speech, and they possess less energy.

Based on emotions, unvoiced speech carries emotion information. For example, if a person is extremely happy or sad or angry, then the possibility of an unvoiced speech is more. In order to catch the contribution of an unvoiced speech, such type of separation is essential.

In a similar manner, silent part or a pause in speech segments varies. For instance, if a person is sad or calm, then the silence part will be a bit longer, whereas if a person is excited or happy or angry, then he/she will speak fast, and there would be reduction in silent part. In order to acquire this information, in this study, we have considered silence duration as a prosodic feature.

When speech utterance was analyzed, it was observed that there are some silence parts where the wave has no energy, or it is extremely small. Unvoiced speech has slightly high amplitude than the silence part, and a voiced speech has high energy. There are several methods to determine if a frame is voiced, unvoiced, or silent. Herein, we have combined two methods, i.e., zero crossing rate (ZCR) and spectrum tilt. ZCR is an indicator of frequency where energy is concentrated in the signal spectrum. Voiced signal exhibits low zero crossing count, unvoiced signal exhibits high zero crossing count, and silence exhibits zero crossing count lower than unvoiced but higher than voiced. The spectrum tilt can be represented by the first-order normalized autocorrelation or first reflection coefficient, which is given by

$$S_t = \frac{\sum_{n=1}^{N} s[n]s[n-1]}{\sum_{n=1}^{N} s^2(n)} \qquad (3)$$

where St is the spectrum tilt, and s[n] and s[n − 1] are consecutive samples. This parameter is reliable as it avoids spike detection in low-level signals. Its ability to detect voiced, unvoiced, and silence is very high.

Figure 1 shows the separation of voiced, unvoiced, and silence frames for a speech signal by combining ZCR and spectrum tilt.

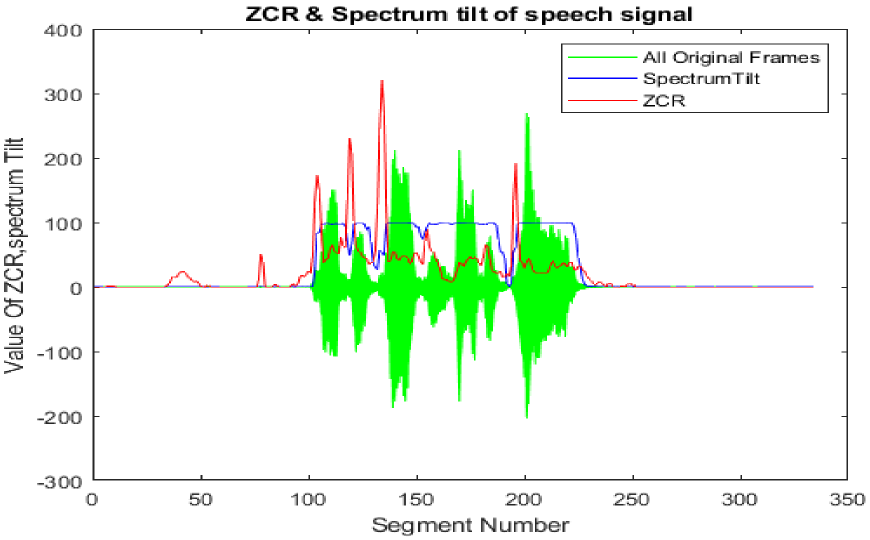

Figure 1: Separation of speech frames into voiced, unvoiced, and silence frames

5.3 System Features

The VT mechanism is stimulated by a series of impulse-like excitations triggered by the stimulation of the vocal folds. A tract structure like this can be thought of as a collection of cavities of differing cross sections. The VT system's sound unit characteristics are treated as the form sequences assumed by the VT when generating separate sound units. A VT functions as a resonator during speech processing, emphasizing specific frequency components depending on the shape of the oral cavity. The VT characteristic is well reflected in the frequency domain analysis of speech signals. Emotion-specific information that is present in the sequence of shapes of the VT may be responsible for producing different sound units in different emotions. This study considered two system features to represent the characteristics of the VT.

5.3.1Fast Fourier Transform Features:

Discrete Fourier transform (DFT) is developed to convert a discrete time signal x[n] to a discrete spectrum by sampling the continuous spectrum X(w) of the signal. Hence, DFT is a powerful computation tool for performing the frequency analysis of discrete time signals. DFT for a given signal can be defined as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{\frac{-j2\pi kn}{N}} \ ; \ k = 0,1 \ldots N-1 \quad (4)$$

where k represents signal harmonics. To avoid an aliasing effect, DFT must be computed for the number of samples in x[n]. Herein, we have separately computed DFT for each voiced and unvoiced speech frame. As higher harmonics do not contain additional information, we have considered first 120 DFT coefficients for each frame. Nevertheless, these 120 DFT coefficients per frame have additional dimensions that lead to heavy computation burden and feature vector overfitting. In order to avoid this, we created a DFT feature vector that consists of min, max, mean, and standard deviation of each harmonic coefficient. We computed these feature vectors for all 120 DFT coefficients for voiced and unvoiced speech frames. Thus, a feature vector for a voiced frame possesses values like 120 min, 120 max, 120 mean, and 120 standard deviation. Overall, there are 480 feature vectors for a voiced speech signal and 480 feature vectors for an unvoiced speech signal. For calculation purposes, FFT algorithms have been used.

5.3.2 MFCC Features:

Mel frequency cepstrum (MFC) can be defined as a short time power spectrum of speech signal, which is calculated as the linear cosine transform of the log power spectrum on a nonlinear mel scale of frequency. MFCCs are coefficients obtained in an MFC representation. In MFC, frequency bands are equally spaced on a mel scale. The mel scale approximates a human’s auditory system response more closely than linearly spaced frequency bands. MFCC depicts the logarithmic perception of loudness and the pitch of a human auditory system. MFCC is widely used for speech emotion recognition [16]. In this study, we have considered 14 MFCC coefficients per frame and have computed MFCC features for voiced and unvoiced speech frames.

5.4 Prosodic Feature:

Prosodic features are voice features derived from long speech samples. During speech processing, humans impose length, modulation, and amplitude patterns on a sound unit chain. By incorporating such prosody restrictions, natural human speech is created. Speech characteristics that are associated with syllables, words, phrases, and sentences are known as prosody. Speech movement tends to be structured by prosody. It usually reflects perceptual speech properties that humans use to perform different speech functions, such as emotion perception [17]. Prosodic characteristics like energy, duration, and tone are thought to be strong emotion correlates [18], [19]. Initial, median, mean, variance, and so on are some of the most important prosodic information sources [20]. In this study, we have considered prosodic features like duration, pitch frequency, and energy.

5.4.1 Duration:

As per the emotion of a speech, the duration of a voiced and an unvoiced speech can vary. Correspondingly, the silent part shall differ. Note that we have considered the ratio of voiced speech duration to unvoiced speech duration as one feature as well as the duration of the silent part as the other feature.

5.4.2 Pitch Frequency:

A speech signal consists of different frequencies that are harmonically related to each other in the form of a series. The lowest frequency of a harmonic series is known as the fundamental frequency or pitch frequency. Pitch frequency is the fundamental vibration frequency of vocal cords, which is in the form of periodic excitation passes via the VT filter and gets convolved with an impulse response of the filter to produce a speech signal. Thus, speech is basically a convolved signal. Fundamental frequency is related to a voiced speech segment. The average male pitchis 100–180 Hz, and for female, it is 160–300 Hz. The fundamental frequency of children is normally from 300 to 500 Hz. There are different methods to compute pitch frequencies. Herein, cepstrum domain method has been used to determine the pitch frequency. As pitch frequency is related to only voiced speech segment, we have computed pitch frequency for all voiced speech frames, and its minimum, maximum, mean, and standard deviation forms a single pitch frequency feature vector. Figure 2(a) shows the pitch frequencies of eight emotions for RAVDESS. Figure 2(b) shows the pitch frequencies of seven emotions for EmoDB. From Figures 2(a) and 2(b), it is evident that as per emotions, pitch frequencies differ.

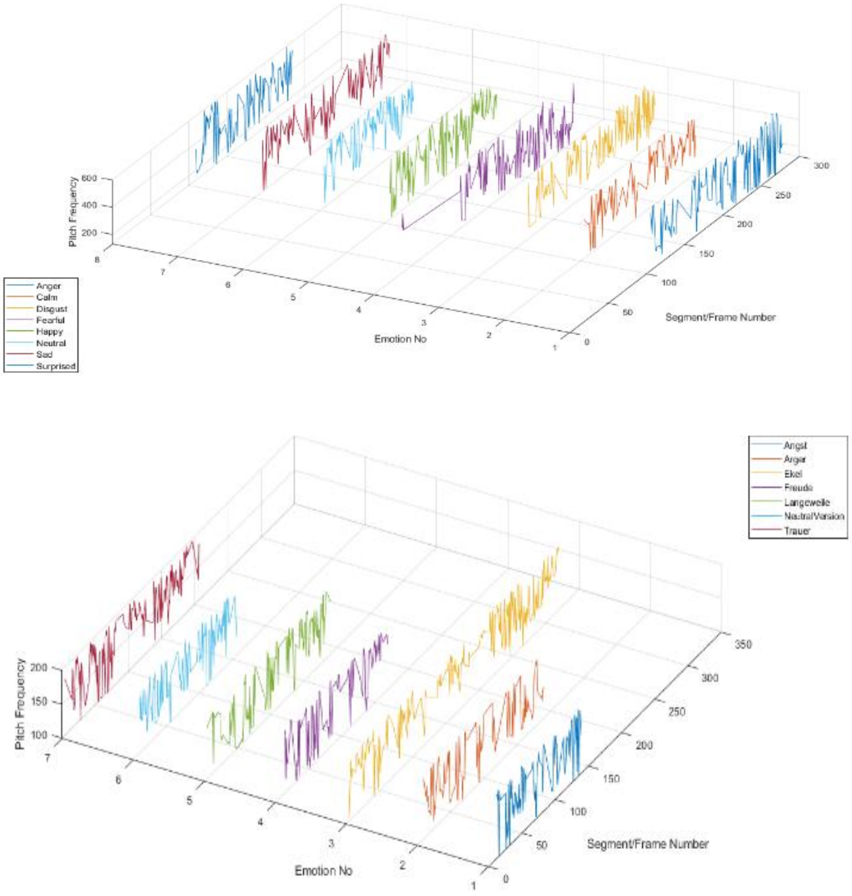

Figure 2. (a) Pitch frequencies of eight emotions for RAVDESS and (b) Pitch frequencies of seven emotions for EmoDB.

5.4.3 Energy:

The energy for each voiced speech segment is calculated as follows:

$$E = \sum_{n=1}^{N} |x[n]|^2 \qquad (5)$$

Herein, with energy for each voiced speech segment, its minimum, maximum, mean, and standard deviation are also considered as a feature. From Figures 3(a) and 3(b), it is evident that there is a difference in energy value as per emotion. The energy peak is highest for disgust and lowest for neutral in RAVDESS.

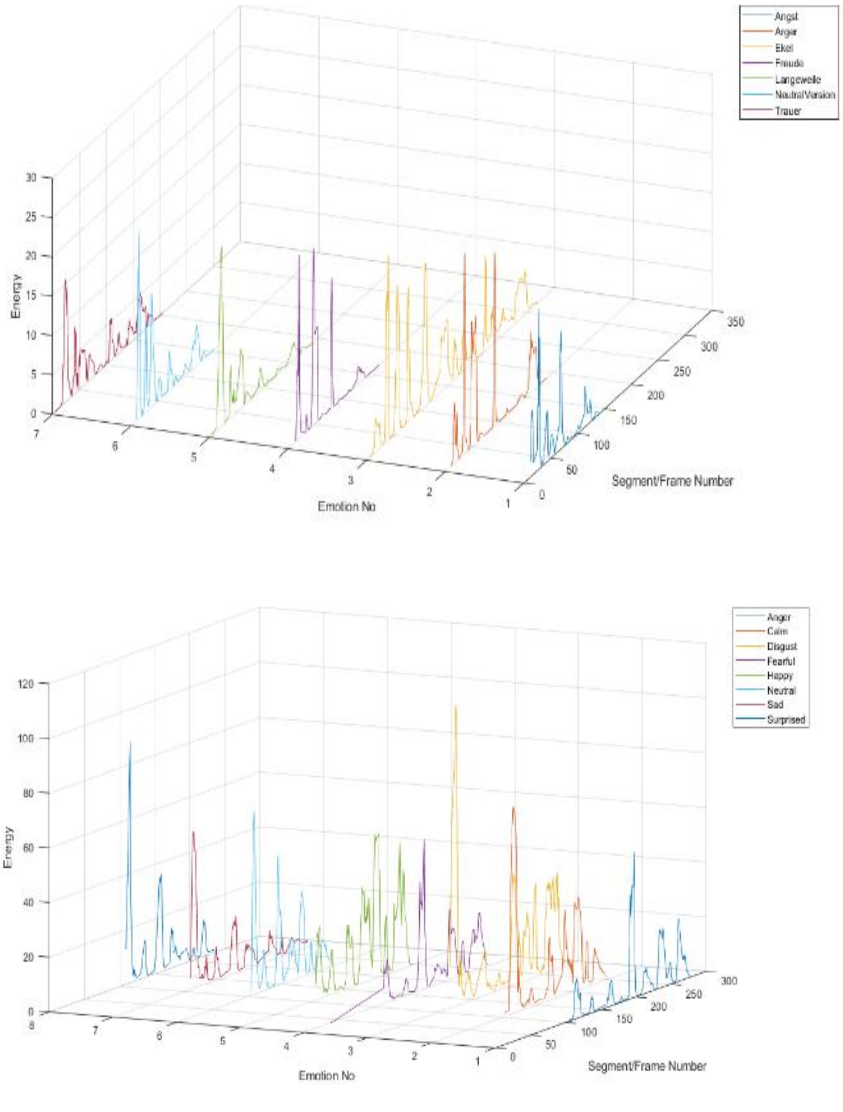

Figure 3. (a) Energy of a voiced speech segment for EmoDB and (b) Energy of a voiced speech segment for RAVDESS.

5.5 Source Features

Source features are speech features that are derived from excitation source signals. Vocal fold vibrations include quasi-periodic impulse-like excitations to VT systems during speech output. Following the withholding of VT characteristics, excitation source signals are derived from speeches.

A linear prediction (LP) residual is the signal that follows. Source functions are features that are derived from LP residuals [12]. The airflow pattern through the glottis, which carries a lot of emotion information, is expressed by GVV signals. The source function that is considered in this analysis is GVV.

5.5.1 GVV:

The glottal pulse signal or GVV is obtained by passing the LP residual signal through a low-pass filter. As excitation is only present for a voiced speech, GVV parameters are extracted only for a voiced speech segment.

The four GVV waveform parameters that were considered in this study as a feature are opening duration, closing duration, opening slope, and closing slope.

From Figures 4(a) and 4(b), it can be observed that GVV shapes are not only different for all emotions but also they are periodic.

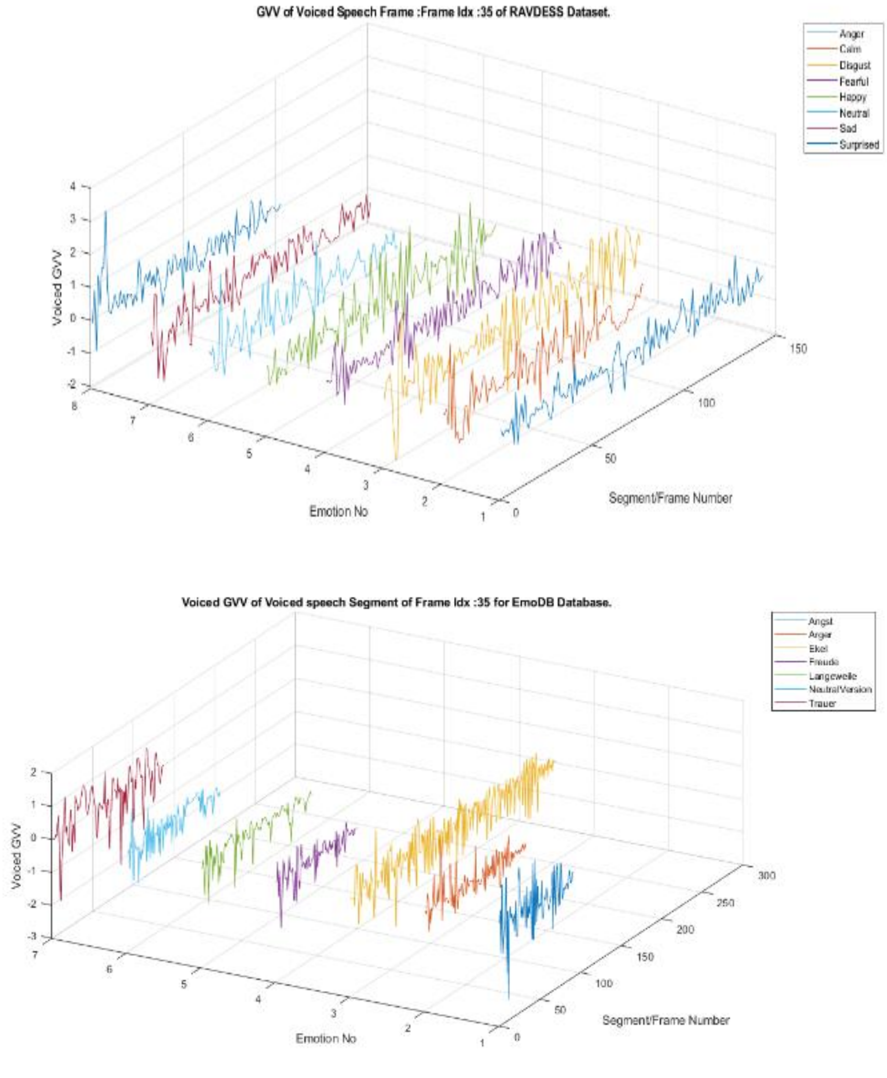

Figure 4: (a) GVV of a voiced speech segment for RAVDEES and (b) GVV of a voiced speech segment for EmoDB.

6. Experimental Results:

In this study, we used only the speech part of both the databases where eight emotions of RAVDESS and six emotions of EmoDB were selected. First, MATLAB® 2019 was used for framing and separating voiced, unvoiced, and silent speech segments. Second, all speech emotion features were extracted. Before feeding the feature vector to the classifier, two issues need to be dealt, i.e., missing data and feature scaling.

6.1 Missing Data:

Based on emotions and the nature of a speaker, the number of voiced and unvoiced speech segments vary, which causes unequal feature vector length for each data. Missing values were normally represented with "NaN". The problem is that most algorithms cannot handle missing values, which needs to be taken care before feeding data to the model. Once the missing values are identified, there are several ways to deal with them.

Either eliminate the features with missing values or add the missing values. By eliminating the features with missing values, there is a possibility of losing essential information about emotions. Therefore, we chose to input the missing values with the mean value of the remaining samples.

6.2 Feature Scaling:

It is a vital step in the preprocessing phase as most machine learning algorithms perform much better when dealing with features that are on the same scale. Feature scaling consists of feature normalization and standardization. The aim is to eliminate speaker and recording variability while ensuring emotional discrimination effectiveness [21]. Herein, we selected feature normalization, which refers to rescaling features to a range from 0 to 1, which is a special case of min–max scaling. For data normalization, min–max scaling method was applied to each feature column.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (6)$$

Once feature vectors were ready, we selected features like MFCC, FFT coefficients, FFT features, ZCR, duration, energy, pitch frequency, and GVV for voiced and unvoiced speech segments. Based on the two databases, SVM classifier was implemented, and the results were compared.

Figure 5 shows the results of using different features for emotion recognition on both the databases. First, we used FFT coefficients with MFCC and prosodic features like energy and duration, which offered an overall accuracy of 68.75% and 47.35% for EmoDB and RAVDESS, respectively.

Thus, instead of using FFT coefficient, we considered system features like voiced FFTF and voiced MFCC in addition to a prosodic feature, which improved the accuracy to 72.92% and 51.39% for EmoDB and RAVDESS, respectively. To observe the effect of unvoiced speech segment features on SER, voiced and unvoiced speech segment FFTF and MFCC features were used, which improved the accuracy to 79.17% and 61.8056% for EmoDB and RAVDESS, respectively.

Finally, along with voiced MFCC, unvoiced MFCC, voiced FFTF, and unvoiced FFTF source features, GVV was given, which offered the highest speech emotion recognition rate. When EmoDB was used for a speaker-independent emotion recognition, the accuracy was 80.85% with 76.93% for fear, 91.67% for anger, 71.43% for happiness, 68.75% for boredom, 87.5% for neutral, and 88.88% for sadness.

On the other hand, for RAVDESS, the accuracy was 70.19% with 73.69% for angry, 63.16% for calm, 73.69% for disgust, 63.16% for fearful, 57.89% for happy, 100% for neutral, 70% for sad, and 60% for surprised. Tables 1 and 2 represent the confusion matrices of FFTF, MFCC, and GVV features on EmoDB and RAVDESS, respectively

Table 1: Confusion Matrix of Emotion Recognition using Voiced, Unvoiced MFCC, FFTF, and GVV Features on EmoDB (%)

	Fear	**Anger**	**Happy**	**Bore**	**Neutral**	**Sad**
Fear	**76.923**	7.6	15.3			
Anger		**91.666**	8.33			
Happy	14.28	14.28	**71.42**			
Bore	6.2			**68.75**	2.5	
Neutral				6.25	**87.5**	6.25
Sad				11.1		**88.88**

Table 2: Confusion matrix of Emotion Recognition using Voiced, Unvoiced MFCC, FFTF, and GVV Features on RAVDESS (%)

	Angry	**Calm**	**Disgust**	**Fearful**	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	**73.6842**		10.5263					15.789
Calm		**63.15**		5.263	5.263		26.31	
Disgust	5.263		**73.6842**		5.263			15.789
Fearful	5.263			**63.1579**	5.263		10.52	15.789

Happy		5.263	5.263	15.789	**57.8**		5.263	10.526
Neutral						**100**		
Sad	5.263	5.263		5.263	10.526	5.263	**70**	
Surprise	10.5263		10.5263	10.5263	10.526			**60**

SER is less for RAVDESS as it consists of eight emotions, which are highly correlated. RAVDESS consists of only two sentences, which may lead to less variation in linguistics comparative to EmoDB, which has ten different sentences.

Neutral gives 100% emotion recognition as in RAVDESS the number of samples for neutral are less and are only in normal intensity.

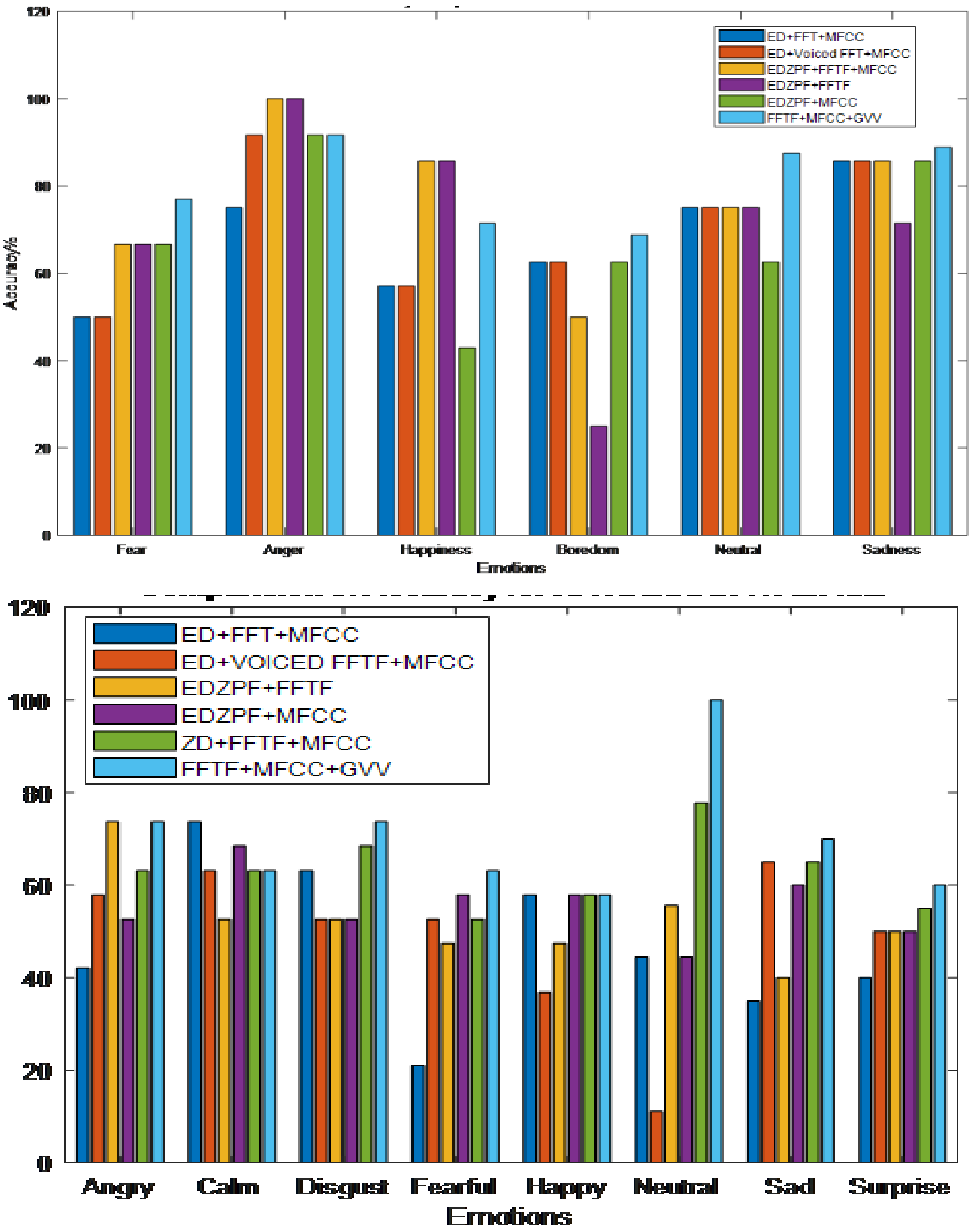

Figure 5. (a) Comparison of accuracy for EmoDB and (b) Comparison of accuracy for RAVDESS

7. Conclusion:

This research suggested FFTF, in which various features were used to identify speech emotion. In addition, it was determined that unvoiced speech segment improves the result. When the system feature MFCC is combined with the features of FFT accompanied by the source feature GVV, it offers highest accuracy for EmoDB as well as RAVDESS. For RAVDESS, a highest accuracy for eight emotions was achieved.

References

[1] S. R. Krothapalli and S. G. Koolagudi, "Characterization and recognition of emotions from speech using excitation source information," Int. J. Speech Technol., vol. 16, no. 2, pp. 181–201, Jun. 2013, DOI: 10.1007/s10772-012-9175-z.

[2] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," Int. J. Speech Technol., vol. 23, no. 1, pp. 45–55, Mar. 2020, DOI: 10.1007/s10772-020-09672-4.

[3] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech." Accessed: Apr. 02, 2021. [Online]. Available: https://github.com/rajamohanharesh/Emotion-Recognition.

[4] S. G. Koolagudi, Y. V. S. Murthy, and S. P. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," Int. J. Speech Technol., vol. 21, no. 1, pp. 167–183, Mar. 2018, DOI: 10.1007/s10772-018-9495-8.

[5] A. Iqbal and K. Barua, "A Real-time Emotion Recognition from Speech using Gradient Boosting," Apr. 2019, DOI: 10.1109/ECACE.2019.8679271.

[6] M. Hao, Y. Tianhao, and Y. Fei, "The SVM based on SMO optimization for Speech Emotion Recognition," in Chinese Control Conference, CCC, Jul. 2019, vol. 2019-July, pp. 7884–7888, DOI: 10.23919/ChiCC.2019.8866463.

[7] M. Ghai, S. Lal, S. Duggal, and S. Manik, "Emotion recognition on speech signals usingmachine learning," in Proceedings of the 2017 International Conference On Big Data Analytics and

[8] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, 2015, doi: 10.1109/TAFFC.2015.2392101.

[9] X. Chen, L. J. Wu, A. Mao, and Z. H. Zhan, "A New learning scheme of emotion recognition from speech by using mean fourier parameters," 11th Int. Conf. Adv. Comput. Intell. ICACI 2019, pp. 96–101, 2019, doi: 10.1109/ICACI.2019.8778548.

[10] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech emotion recognition: Methods and cases study," ICAART 2018 - Proc. 10th Int. Conf. Agents Artif. Intell., vol. 2, no. March 2019, pp. 175–182, 2018, doi: 10.5220/0006611601750182.

[11] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, 2012, doi: 10.1007/s10772-011-9125-1.

[12] C. Hsu and C. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," vol. 13, no. 2, pp. 415–425, 2002.

[13] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Commun., vol. 116, no. October 2019, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.

[14] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011, doi:

10.1016/j.patcog.2010.09.020.

[15] K. S. Rao, B. Yegnanarayana, and S. Member, “Prosody modification using Instants of Significant Excitation,” vol. 14, no. 3, pp. 972–980, 2006.

[16] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, “Emotion recognition by speech signals,” EUROSPEECH 2003 - 8th Eur. Conf. Speech Commun. Technol., pp. 125–128, 2003.

[17] T. L. Nwe, S. Wei, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” vol. 41, pp. 603–623, 2003, doi: 10.1016/S0167-6393(03)00099-2.

[18] M. Schr, “Emotional Speech Synthesis : A Review,” pp. 2–5, 2001.

[19] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, “Iterative feature normalization scheme for automatic emotion detection from speech,” IEEE Trans. Affect. Comput., vol. 4, no. 4, pp. 386–397, 2013, doi: 10.1109/T-AFFC.2013.26.

[20] S. Patil and G. K. Kharate, "A Review on Emotional Speech Recognition: Resources, Features, and Classifiers," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 669-674, doi: 10.1109/ICCCA49541.2020.9250765.

[21] P. D. Das and P. S. Sengupta, “An Emotion Based Speech Analysis,” vol. 4, no. 10, pp. 295–302, 2015.

[22] J. Bachorowski, “Vocal Expression and Perception of Emotion,” pp. 53–57, 1999.

[23] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391]