MFCC and Machine Learning Based Speech Emotion Recognition on TESS and IEMOCAP Datasets

Muhammad Zafar Iqbal[1], Ghazanfar Farooq Siddiqui[1]

[1]Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Federal, Pakistan

Corresponding author: Muhammad Zafar Iqbal (mziqbal@ cs.qau.edu.pk)

Abstract: Emotions in speech provide a lot of information about the speaker's emotional state. This paper presents a classification of emotions using a support vector machine (SVM) with Mel Frequency Cepstrum Coefficient (MFCC) features extracted from the voice signal. We have considered the following five emotions, namely anger, happy, neutral, pleasant surprise and sadness, for classification purposes. The proposed methodology, including SVM-Gaussian and SVM-Quadratic, is tested for its performance on the Toronto Emotion Speech Set (TESS) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. Our proposed methodology achieved 97% accuracy with TESS and 86% with IEMOCAP datasets, respectively.

Keywords: Emotion Recognition, Machine Learning, MFCC, SVM, TESS, IEMOCAP.

1 Introduction

The audio speech signal is the fastest and most natural means of communication between humans. This fact prompted researchers and scientists to use the speech signal as a means of human-machine interaction to make the machine more efficient [1,2]. SER (Speech Emotion Recognition) is one of the most commonly used methods to improvise efficiency and standardize human-machine interaction and thus improve artificial intelligence. In SER, the speech signal articulates different types of emotions such as anger, neutral speech, sadness, happiness, excitement, fear and many more. The voice signal is processed on a paralinguistic basis [3]. SER has many applications; it can be used in the on-board vehicle system to identify the driver's mental state for his safety. Moreover, it can be used in call centers to detect customers' emotional behavior [4]. Much research has been done in SER, but the problem of accuracy is still there. In this work, we used two databases for emotion recognition. The Speech signal illustration is given in Figure 2.

1.1 TESS Dataset

This corpus comprises a set of 200 target words spoken by two English actresses aged 26 and 64. The dataset depicts seven emotions (happiness, anger, disgust, fear, pleasant surprise, neutral and sadness) and consists of 2800 audio files in total. TESS only has two speakers of different ages, which are treated as separate databases, one speaker per database. Since the TESS database just had one speaker for the youngest and oldest database, speaker-specific information could not affect accuracy. In this work, we considered four classes of emotions (angry, happy, disgust and pleasant surprise) for prediction purposes [5].

1.2 IEMOCAP Dataset

This database [6] is multimodal and multi-speaker, recently collected at USC's SAIL Lab. We have included three emotions (angry, neutral, and sad) for this purpose. Speech is the expression of the human voice through a computing device that recognizes emotions. This means that computational software containing mathematical algorithms capable of calculating the emotions of our speech through a series of tools. Speech recognition research enables the machine to understand the speaker's emotions and use this information during human-machine interaction. Speech includes information such as speed, tone of voice, content, etc., with which emotions play an essential role in human interaction.

1.3 SVM Classifier

The SVM is a widely used classifier due to its accuracy and ability to process large data [7]. It is also one of the best machine learning methods that provide better results. SVM is a statistical classifier that classifies data into multi classes or binary classes based on training data. SVM is quite suitable if the size of the dataset is not too large. The high dimensional spaces are used to build the high dimensional space that can be used for regression, classification, or other related tasks [8]. For speech emotion training and testing, features are extracted and then models are created. SVM helps distinguish between multi-class datasets by creating a hyperplane and then indicating which class it belongs to. The dataset of multiple classes are used as an input together that simultaneously map the dataset with the suitable class label to help the prediction of the best results to train the SVM model. When testing a dataset, this hyperplane maps the testing data into the training data portion closest to the class.

2 Related Work

One of the essential steps in recognizing vocal emotions is feature extraction. The extraction of speech characteristics in the categorization problem involves reducing the dimensionality of the input vector space while maintaining the distinguish power of signals. Researchers have come up with many essential speech emotion features as well as speech emotion feature extraction methods in recent years. Dujaili et al. [9] performed a classification of emotions using K-Nearest Neighbor (KNN) and SVM. The total number of extracted features is reduced through principal component analysis.

This research [10] proposes to combine auditory and textual information by applying a late fusion approach in two steps. First, the auditory and textual features are trained separately in deep learning systems. Second, the deep learning systems' prediction results are fed into SVM to predict the final regression score. Also, this research's task is the dimensional modelling of emotions, as it can allow a more in-depth analysis of effective states. Abdul et al. [11] proposed a face expression based emotion recognition method using SVM to improve the detection performance with multiple emotions effectively. The average obtained results are much better than other existing techniques.

3 Proposed Methodology

Speech Emotion's recognition work consists of two steps, MFCC feature extraction from speech signal and classification of emotions using SVM classifier. We used the MFCC of the speech signals as the extracted features; then, we performed classification on these extracted features using the SVM classifier to identify the voice emotions. The computation of the Mel Frequency Cepstral coefficients involves the following steps with a brief description of the task to extract the speech characteristics. Our proposed methodology is given in Figure 1.

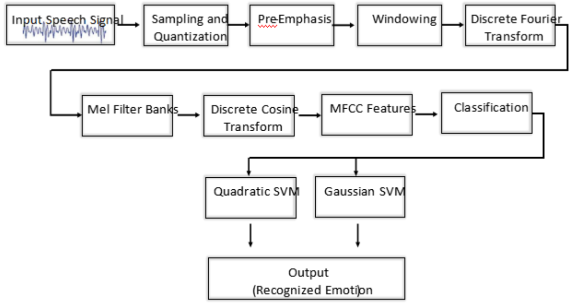

Figure 1: **Proposed methodology for speech emotion system using Quadratic SVM and Gaussian SVM, with various preprocessing steps**

3.1 Sampling and Quantization

As we know that the audio wave is a continuous signal while the computer is a digital machine that cannot directly represent the continuous signals, so we have to convert these continuous signals into a discrete finite set of information; this is called sampling. While quantization is the process of representing real-valued numbers as a set of integers [12].

3.2 Pre-emphasis

The spectrum of speech has higher energy at low frequencies compared to high frequency [1]. Through pre-emphasis, energies are increased to high-frequency levels, thereby balancing the level of energies in the spectrum. Pre-emphasis is considered a noise reduction tool because it reduces the power of the noise without affecting the rest of the signal [13].

3.3 Windowing

The speech signal varies over time but is stationary for short segments. So in this step, the speech signal is split into short chunks of rectangular segments of about 10ms for further analysis by minimizing signal discontinuities using the Hamming window, which reduces the signal's amplitude towards zero at the edges of the window thus avoiding discontinuity. Below are plots of the windowing and hammering window effect [14].

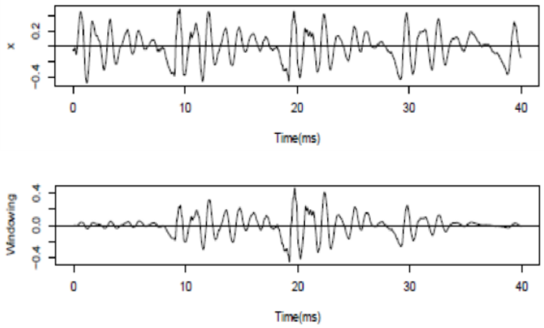

Figure 2: **Speech signal illustration**

3.4 Discrete Fourier Transform

After windowing, the next step is DFT, and this step is based on the spectral analysis to extract the speech features based on the magnitude spectrum calculation [15]. At this point, the energy level/spectral information of each window is collected. The input to DCT is the windowed signal, and the output is the complex numbers representing the amplitude and phase of the frequency.

The commonly used algorithm to discover DCT is the Fast Fourier Transform (FFT).

3.5 Mel Filter Bank

Refer to the human capacity audit, which we are most sensitive to between 20 and 1000Hz. The Mel frequency spectrum calculation is done after the DFT by passing the spectrum through the Mel filters to obtain Mel Cepstrum. It is used to match the human ear's frequency resolution by linearly spacing the energies of frequencies below 1000Hz and logarithmically spacing them after that. Figure 3 represents the Mel frequency filter bank with 12 bins.

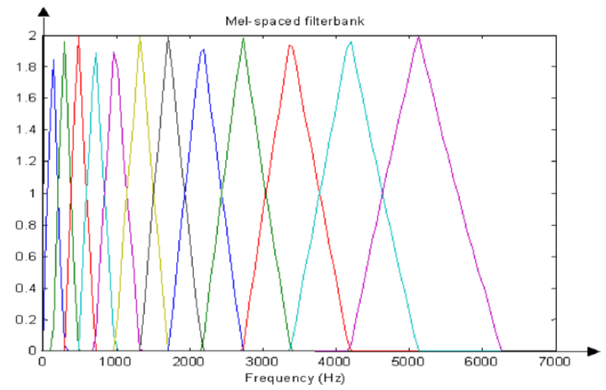

Figure 3: **Mel frequency filter bank with 12 bins**

3.6 Extracting MFCC Features

As we know, feature extraction refers to the method used to reduce the dimensionality of the given dataset and hence minimize system time and space. The feature vector contains discriminant information that distinguishes one object/instance from another object/instance [13], so it is essential to select an appropriate feature extraction algorithm. In this article, we have used MFCCs as extracted characteristics for further classification of emotions. MFCC is the classic, efficient and successful approach used for speech-related tasks such as gender identification by voice, speech recognition, speech emotion recognition and many more. MFCC coefficients are successful because they are derived from human speech patterns [16]. The MFCC attempts to mimic the human ear, where the audio frequency is determined as an Asymmetric Spectrum [17]. Mel's scale is the scale of perception of pitches that the audience considers equal to each other's distance. Mel comes from the word melody, which indicates that the scale is based on pitch comparisons. The reference point between this scale and the normal frequency measurement is defined as a tone of 1000Hz, equal to 40dB above the listener's threshold, with a height of 1000 mels. The mel scale corresponds to linearly below 1kHz and logarithmically above 1kHz [18].

Table 1: **Performance comparison of our technique with other state-of-the-art techniques on TESS dataset.**

Authors	**Year**	**Accuracy**
Dupuis et al. [19]	2011	82 %
Praseetha et al. [20]	2018	95.82 %
Huang et al. [21]	2019	85 %
Our Method	**2020**	**97 %**

3.7 Classification

After feature extraction, the next step is to categorize emotions. In this work, we have used the following two types of SVM for the Toronto speech dataset and the IEMOCAP dataset, respectively:

Gaussian Support Vector Machine and Quadratic Support Vector Machine.

4 Experiments and Result Discussion

As mentioned earlier, in this work, we experimented with SER on two databases: TESS and IEMOCAP. In TESS, we have divided the dataset into training and testing parts. For training purposes, 200 samples (50 samples from each category) of emotions (angry, happy, disgust, pleasant surprise) were separated. At the same time, 24 samples (6 from each category) were separated from testing the data. The MFCC vector of this training data was passed to the classifier, which gave 97% accuracy. In the IEMOCAP speech dataset, 150 samples (50 for each category) of emotions (angry, neutral, and sad) were passed to the classifier as a training dataset. At the same time, 30 samples (10 from each category) were treated as test data. The accuracy of this dataset after training was 86%. Table 1 shows the performance comparison of our technique with three other state-of-the-art techniques. The results in Table 1 show that our proposed method outperforms other three baseline methods.

5 Conclusion and Future Work

In this paper, we have tested our methodology on two databases with two different types of SVM. In total, five distinct emotions have been classified. The results are obtained by placing cepstral coefficients in the feature vector. Where each row represented each signal, and the matrix was formed by adding labels with an interval of 50 rows (one label for 50 samples of each category). This work can be improved by adding preprocessing steps before feature extraction and considering more artifact features (like pitch, time domain, etc.) as well as current features. Moreover, in addition to these databases, other popular databases such as Berlin Speech Database can be added to improve accuracy.

References

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

[2] L. Hussain, I. Shafi, S. Saeed, A. Abbas, I. A. Awan, S. A. Nadeem, and B. Rahman, “A radial base neural network approach for emotion recognition in human speech,” *IJCSNS*, vol. 17 , no. 8, p. 52, 2017.

[3] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Fifteenth annual conference of the international speech communication association*, 2014.

[4] Y. Pan, P. Shen, and L. Shen, “Speech emotion recognition using support vector machine,” *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

[5] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (TESS),” 2020.

[6] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[7] K. Aurangzeb, N. Ayub, and M. Alhussein, “Aspect based multi-labeling using svm based ensembler,” *IEEE Access*, vol. 9, pp. 26026–26040, 2021.

[8] B. Jena, A. Mohanty, and S. K. Mohanty, “Gender recognition of speech signal using knn and svm,” *Available at SSRN 3769786*, 2021.

[9] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, “Speech emotion recognition based on svm and knn classifications fusion,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, p. 1259, 2021.

[10] B. T. Atmaja and M. Akagi, “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm,” *Speech Communication*, vol. 126, pp. 9–21, 2021.

[11] M. H. Abdul-Hadi and J. Waleed, “Human speech and facial emotion recognition technique using svm,” in *2020 International Conference on Computer Science and Software Engineering (CSASE)*. IEEE, 2020, pp. 191–196.

[12] A. Host-Madsen and P. Handel, “Effects of sampling and quantization on single-tone frequency estimation,” *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 650–662, 2000.

[13] T. Seehapoch and S. Wongthanavasu, “Speech emotion recognition using support

vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*.
IEEE, 2013, pp. 86–91.

[14] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7527–7531.

[15] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *2017 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2017, pp. 583–588.

[16] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4821–4824.

[17] A. Charisma, M. R. Hidayat, and Y. B. Zainal, "Speaker recognition using mel-frequency cepstrum coefficients and sum square error," in *2017 3rd International Conference on Wireless and Telematics (ICWT)*. IEEE, 2017, pp. 160–163.

[18] E. Wong and S. Sridharan, "Comparison of linear prediction cepstrum coefficients and melfrequency cepstrum coefficients for language identification," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*. IEEE, 2001, pp. 95–98.

[19] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[20] V. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *Journal of Computer Science*, vol. 14, no. 11, pp. 1577–1587, 2018.

[21] A. Huang and P. Bao, "Human vocal sentiment analysis," *arXiv preprint arXiv:1905.08632*, 2019.