Rhythm Formant Analysis for Automatic Depression Classification

Kumar Kaustubh[1], Parismita Gogoi[2,3], and S.R.M Prasanna[1]

[1] Indian Institute of Technology, Dharwad, India
{221022003, prasanna}@iitdh.ac.in
[2] Indian Institute of Technology, Guwahati, India
[3] DUIET, Dibrugarh University
parismitagogoi@iitg.ac.in

Abstract. This paper presents a study on the application of Rhythm Formant Analysis (RFA) for automatic depression classification in speech signals. The research utilizes the EATD-corpus, a dataset specifically designed for studying depression in speech. The goal is to develop an effective classification system capable of distinguishing between depressed and non-depressed speech based on Rhythm Formant (RF) features. The proposed methodology involves extracting RFs from the speech signals using signal processing techniques. Two kinds of RFs, namely Amplitude Modulation (AM) and Frequency Modulation (FM) RFs and their combinations are analyzed and used as features for classification. These features provide valuable information about the temporal and spectral characteristics of the speech. The classification system is built using a Decision Tree (DT) classifier and its results are compared with logistic regression and random forest. The model's performance is evaluated using the accuracy, F1 scores for each class and their macro and weighted averages. Experimental results demonstrate promising outcomes, with the DT classifier achieving an accuracy of 70%, a weighted average F1 score of 0.73 and a macro average F1 score of 0.53 when using FM RFs as feature, showing much better performance compared to other features and classifiers. These results indicate that the proposed approach effectively captures discriminative features related to depression in the speech signals. The findings suggest that RFs have the potential to serve as a valuable tool for building automatic depression classification systems.

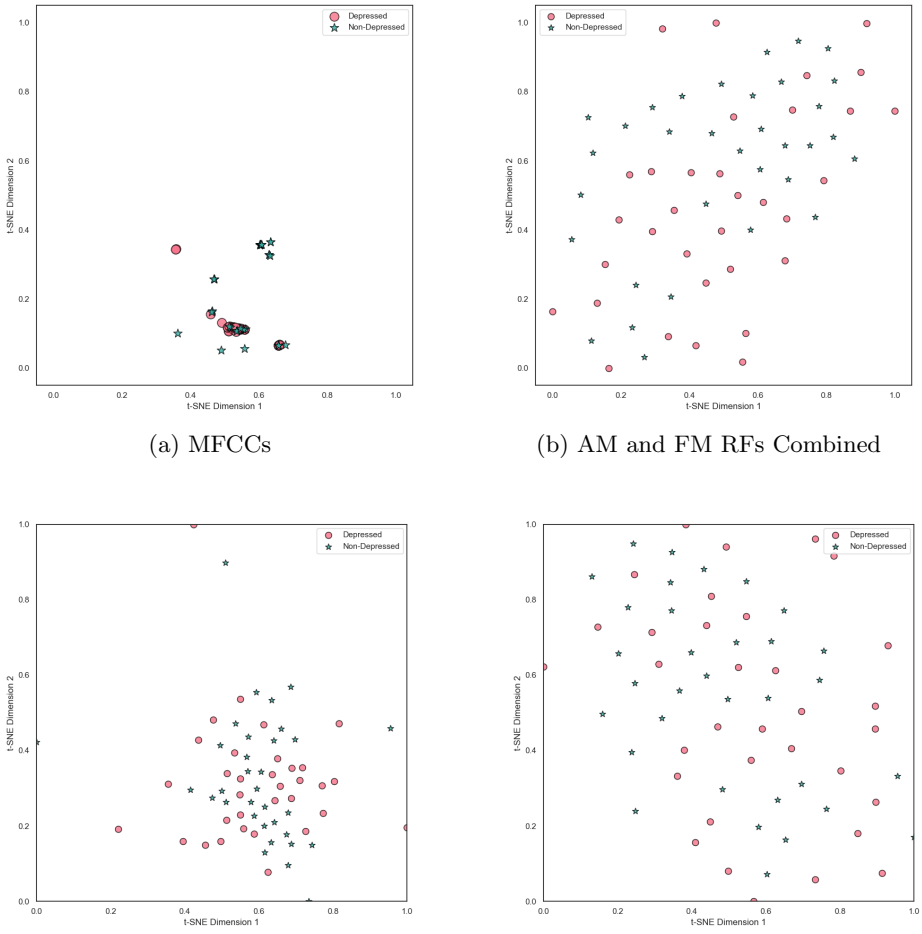
Keywords: Rhythm Formant Analysis· Depression Classification· EATD-corpus.

1 Introduction

Depression, a mental health disorder affecting millions worldwide, poses a significant challenge to clinicians and researchers alike due to its complex nature and varying symptoms. Accurate and timely diagnosis of depression is crucial for effective treatment and intervention. Depression is characterized by persistent feelings of sadness, hopelessness, and a loss of interest in daily activities.

Clinicians currently rely on a variety of methods to diagnose depression, including interviews, self-report questionnaires, and psychiatric assessments [2]. While these approaches have proven valuable in many cases, they are subjective and prone to biases. The interpretation of interview responses and questionnaire results can vary based on the clinician's experience, cultural background, and personal biases, subsequently leading to misdiagnosis or delayed treatment. Speech, as a potential candidate for depression detection, offers several advantages over traditional diagnostic methods. Speech is a fundamental aspect of human communication and reflects various emotional and cognitive states [3]. In the past, researchers have come up with many acoustic features which were capable of capturing the patterns in the speech of a depressed individual. One such study involves the analysis of GMM based features along with power spectral densities within multiple sub-bands which turned out to be effective in addressing patterns associated with depressed and non-depressed speech samples [11]. While these features have been helpful in such tasks, deep neural networks in combination with hand-crafted features have also been utilized in order to achieve better performance [7]. Using deep learning, attempts have been made to develop systems for measuring depression severity by combining transfer learning, attention based learning and unsupervised learning [12]. Researchers have also implemented deep learning architectures involving convolutional and recurrent neural networks using spectrograms for similar emotion recognition tasks [9].

In this paper, we aim to explore the potential of Rhythm Formant Analysis (RFA) as a novel approach for automatic depression classification. By investigating the relationship between speech rhythm patterns and depressive states, we seek to contribute to the ongoing efforts in developing objective and reliable tools for depression assessment. One of the relevant features used for classifying speeches of depressed and non-depressed subjects are Mel Frequency Cepstral Coefficients (MFCCs) [1], which provide compact representation of the speech signals and are effective in capturing important characteristics of speech. Rhythm Formants (RFs), on the other hand, have been employed by researchers in machine learning to classify different speaking styles in speech-related classification tasks [4].This serves as one of the motivations to investigate the distinctive rhythm patterns that might be associated with speeches of depressed and non-depressed individuals. Based on our initial investigation, we have observed notable variations in the arrangement of t-SNE plots for MFCCs, Amplitude Modulation (AM), Frequency Modulations (FM) RFs and their combination, as shown in Fig.1 below. Fig.1 (a) is the 2-dimensional t-SNE representation of 13-dimensional MFCCs, Fig.1 (b) shows the 2-dimensional t-SNE representation after concatenation of 6-dimensional AM RFs and 6-dimensional FM RFs to create a 12-dimensional feature space. Fig.1 (c) and Fig.1 (d) presents the 2-dimensional t-SNE plots for 6-dimensional AM and FM RFs, respectively. These variations indicate that these features capture different informations which further motivates the usage of RFs for carrying out this classification task. From the t-SNE plots in Fig.1, we can observe that MFCCs are much more confined in the feature space compared to other features.



(a) MFCCs

(b) AM and FM RFs Combined

(c) AM RFs Only

(d) FM RFs Only

Fig. 1: 2-dimensional t-SNE visualization for (a) 13-dimensional MFCCs, (b) 12-dimensional RFs after concatenating 6-dimensional AM and FM RFs, (c) 6-dimensional AM RFs and (d) 6-dimensional FM RFs

The paper is arranged in the following sections: In section 2, AM and FM RFs and their potential use in depression classification have been discussed along with the signal processing steps involved in their extraction. Section 3 presents the classification system built using FM RFs as features with a Decision Tree (DT) classifier. Section 4 describes the dataset and the experiments performed, followed by section 5 where the evaluation metrics and classification results are discussed. Finally, in section 6, the paper is concluded with possible future work directions.

2 Rhythm Formant Analysis

Four assertions in RFA including speech modulation knowledge, simultaneous RFs, serial RFs and asymmetrical rhythms are put forward by Gibbon in his exploratory paper on RFA [5]. The key feature of RFA is the concept of spectral peak values in Low-Frequency (LF) spectrum which are coined as LF rhythm formants. RFA approach explicitly relates to the time-stamps of linguistic categories (e.g. syllable, foot) for each language. Speech rhythms are described as waves with frequencies below 10 Hz [5]. This rhythm can be found by analyzing the changes in loudness and pitch of speech at these LFs. RFA involves looking at the variations in pitch (FM) and loudness (AM) across the whole sample. In our study, we employ a method called RF detection, which involves analyzing the spectral peaks of demodulated AM and FM signal envelopes in the frequency domain [6]. This approach allows us to identify and characterize rhythmic patterns present in the speech signals. For extracting the FM RFs, we first normalize the speech signal between -1 and 1. Next, we obtain the pitch contour and transform it into the frequency domain using Fourier transform. Then, we extract the LF segments (0 Hz to 10 Hz) from the magnitude of the Fourier transform and normalize it to have a value between 0 and 1. Finally, we pass it through a peak-picking algorithm to obtain the FM RFs [6]. Fig.2 illustrates these steps using a block diagram.

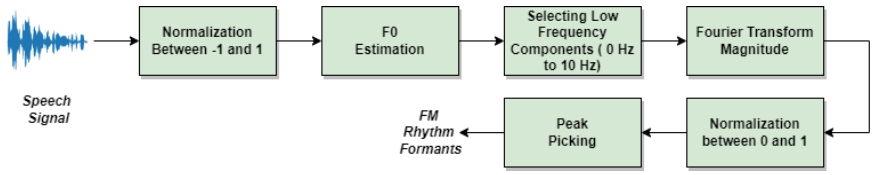


Fig. 2: Block diagram showing signal processing steps for extraction of FM rhythm formants

Extraction of AM RFs involves a similar approach where we first normalize the speech signal between -1 and 1 and obtain the amplitude envelop using the absolute value of Hilbert transform. After passing the amplitude envelope through a median filter, we then transform the resulting signal into frequency domain using Fourier transform. Again, we extract the LF components (0 Hz to 10 Hz) from the magnitude of the Fourier transform and normalize it between 0 and 1. Finally, we use a peak picking algorithm to obtain the AM RFs [6]. The same is illustrated in the block diagram shown in Fig.3. Using the FM RF extraction steps, the generated FM envelope and peaks in the LF spectrum, which represent the FM RFs, are shown in the Fig.4 below for a depressed and a non-depressed speech sample. Similarly, using AM RF extraction steps, the generated AM envelope and peaks in the LF spectrum, which represent the AM RFs are shown in Fig.5 below for the same depressed and non-depressed speech samples

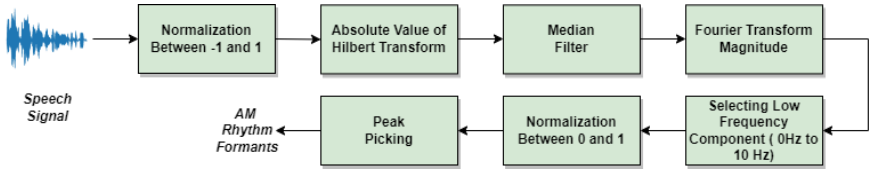


Fig. 3: Block diagram showing signal processing steps for extraction of AM rhythm formants

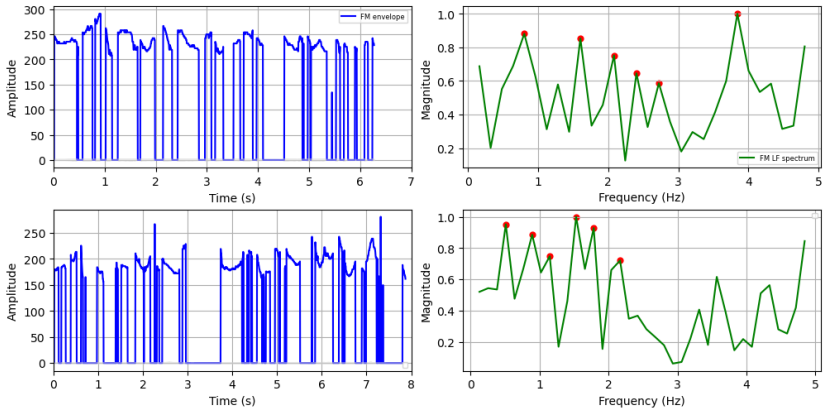


Fig. 4: FM envelope and LF spectrum of a non-depressed (top panel) and depressed speech sample (bottom panel). The peaks in the LF spectrum are representing FM RFs

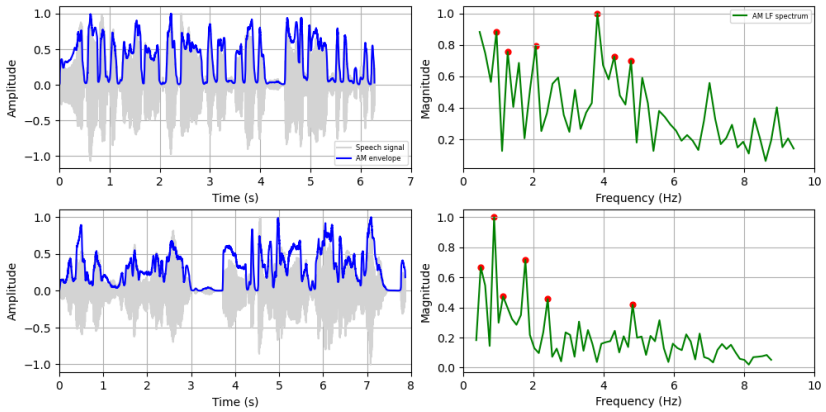
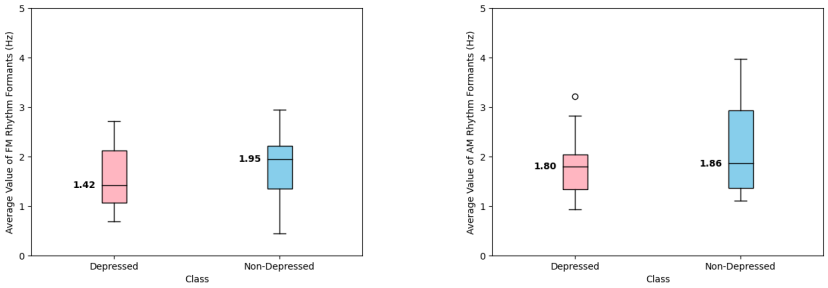


Fig. 5: AM envelope and LF spectrum of a non-depressed (top panel) and a depressed speech sample (bottom panel). The peaks in the LF spectrum are representing AM RFs



(a) FM rhythm formants (b) AM rhythm formants

Fig. 6: Box plot showing median values of (a) FM RFs and (b) AM RFs averaged across each of the selected speech samples for depressed and non-depressed class

as used in FM RFs. It can be observed from Fig.4 and Fig.5 that, on average, RFs in case of depressed speech samples are at lower frequencies compared to non-depressed speech samples. The average of FM RFs comes out to be 1.34 Hz and 2.24 Hz for depressed and non-depressed speech samples,respectively, used in the illustration whereas the average in case of AM RFs comes out to be 1.93 Hz and 2.87 Hz for depressed and non-depressed speech samples, respectively.

Strengthening the case for the potential of AM and FM RFs in the classification task, we present a box plot in Fig.6. For this plot, we randomly selected 15 speech samples from each class and extracted the first 6 AM and FM RFs. Next, we calculated the average value for each speech sample across the first 6 AM and FM RFs. After calculating this average, the median values of FM RFs for the depressed and non-depressed class for the selected samples are calculated as 1.42 Hz and 1.95 Hz, respectively as shown in Fig.6 (a). Similarly, the median value of depressed and non-depressed class for these selected samples comes out to be 1.80 Hz and 1.86 Hz respectively for average value of AM RFs across each sample. The lower median value for the depressed class in both the cases suggests that, on average, the AM and FM RFs tend to be lower in individuals classified as depressed. This implies a potential characteristic or trend in the speech patterns of individuals with depression, as compared to those without.

3 Classification System Using FM Rhythm Formants

The paper presents a system for depression classification based on the FM RF feature extraction method. The system utilizes the top 6 FM RF features extracted using signal processing techniques as described in the previous section and uses them as the 6-dimensional feature vector for the classification task. The system employs a standard decision tree classifier which takes in the feature vector to classify the speech samples into two categories: depressed and non-depressed. Although we attempted to build the system by utilizing a combination of AM and FM RFs along with the 13 MFCCs and explored other

classifiers as well, it turns out that using FM RFs alone with a decision tree outperforms the combination. The experiments, detailed results and possible explanations are discussed in the subsequent sections. This system holds the potential to assist mental health professionals in detecting and intervening with individuals at risk of depression, complementing their expertise.

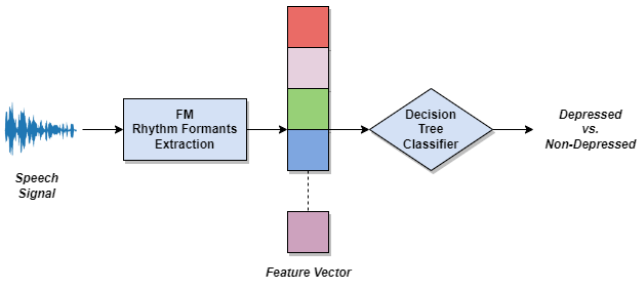


Fig. 7: Classification system built after extracting FM rhythm formants, using them for creating a 6-dimensional feature vector and passing it through a decision tree classifier to obtain a label either as depressed or non-depressed

4 Dataset and Experimental Description

The study employed a Chinese dataset known as Emotional Audio-Textual Depression corpus (EATD-corpus), which consists of speech recordings from individuals classified as either depressed or non-depressed [10]. The dataset includes a total of 162 different speakers, with 30 classified as depressed and 132 as non-depressed. The categorization is based on the indexed Self-Rating Depression Scale (SDS) score (Raw SDS score multiplied by 1.25), where a score of 53 or higher signifies depressed, while a score below 53 indicates non-depressed. The total duration of the dataset is around 2.26 hours. The training dataset consists of 83 speakers, with 19 categorized as depressed and 64 as non-depressed. Each individual provided three responses, leading to a total of 57 depressed samples and 192 non-depressed samples. Likewise, the testing dataset comprises 79 speakers, with 11 identified as depressed and 68 as non-depressed. Each individual in the testing set also provided three responses, resulting in a total of 33 depressed samples and 204 non-depressed samples. Table 1 below shows exact distribution of the dataset.

In our study, we have conducted a detailed experiment that looks into the process of analyzing depressed and non-depressed speech signals. The main goal of this analysis is to extract and examine RFs from the speech signals. Initially, each of the audio files, which are present in WAV format, are loaded using the default sampling rate of 16000 Hz. After the successful import of audio data, signal normalization is performed to ensure that the audio signal fits within a

Table 1: Distribution of the dataset

	Depressed	**Non-Depressed**	**Total Data**
Train	57 samples (14 min 06 sec)	191 samples (56 min 07 sec)	70 min 13 sec
Test	33 samples (7 min 48 sec)	204 samples (50 min 28 sec)	58 min 16 sec

standardized range of -1 to 1, achieved by dividing the signal by its maximum value. For the extraction of FM RF features, the fundamental frequency (F0) is estimated using the AMDF (Absolute Magnitude Difference Function) algorithm, which calculates the difference between a particular frame of a signal and its delayed version which results in smallest sum of absolute difference. After F0 estimation, we select the LF components between 0 Hz to 10 Hz, and a Fourier transform is computed. The magnitude of the Fourier transform is then normalized between 0 and 1. Finally, this normalized magnitude is passed through a peak picking algorithm to detect the top 6 peaks, which are the FM RFs. For the calculation of AM RFs, a similar procedure is followed. After normalizing the speech sample, the envelope for AM signal is calculated. The Hilbert transform of the signal is computed to gain the instantaneous phase and amplitude, and then the absolute value of the Hilbert transform is derived to ascertain the envelope. To get a smoother signal, a median filter, with a window size of 501, is applied to the envelope, which is then normalized by dividing it by its maximum value. Subsequently, a spectral analysis is performed on the normalized envelope, and the LF segment of 0 Hz to 10 Hz is extracted from the magnitude component of the Fourier transform of the envelope. After extraction, the magnitude components are normalized between 0 and 1. Finally, this normalized magnitude is passed through a peak picking algorithm to detect the top 6 peaks, which are the AM RFs. We have also extracted 13 MFCCs and used them as feature vector with different classifiers for comparing with the RF results.

After extracting these features, they form the basis of the four separate experiments conducted in this study. In the first experiment, we utilize only the 13 MFCCs as the 13-dimensional feature vector. The second and third experiment focuses solely on AM and FM RFs. Using both AM and FM RFs separately, we generate 6-dimensional feature vector for each case. In the fourth and final experiment, a combination of AM and FM RFs is used by concatenating them and creating a 12-dimensional feature vector. All the four feature vectors are then used in three different machine learning algorithms: DT classifier, random forest and Logistic Regression (LR) to carry out the binary classification task, which finally provides us with the label either as depressed or non-depressed. The results of these experiments along with the evaluation metrics are described in the next section.

5 Evaluation Metrics and Results

In the conducted study, we aimed to evaluate the performance of various machine learning classifiers for detecting depression based on the experiments conducted.

Following key evaluation metrics were used to assess the performance of the models: accuracy, F1 scores of each class, Macro Average (MA) F1 score, and Weighted Average (WA) F1 score. We have also tabulated the values of Precision (P) and Recall (R) for each class as well as their MA and WA. Apart from these evaluation metrics, we also present the confusion matrices for the four experiments conducted for the DT classifier. Accuracy (A) measures the percentage of correct predictions made by the model [8]. It is given by the formula:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP represents the number of True Positive predictions, TN represents the number of True Negative predictions, FP represents the number of False Positive predictions, and FN represents the number of False Negative predictions. Precision, on the other hand, is a measure of the accuracy of positive predictions. It calculates the proportion of correctly predicted positive instances (TP) out of all instances predicted as positive (TP + FP). Precision focuses on the correctness of positive predictions and provides insights into the system's ability to minimize FPs [8]. The Precision (P) is given by the following formula:

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as TP rate, measures the system's ability to correctly identify positive instances. It calculates the proportion of TP predictions out of all actual positive instances (TP + FN). Recall emphasizes the completeness of positive predictions and provides insights into the system's ability to minimize FNs [8]. Recall (R) can be calculated as follows:

$$R = \frac{TP}{TP + FN} \tag{3}$$

The F-Measure or F1 score is the harmonic mean of P and R [8], where P and R are precision and recall as described above. The F1 score is given by the formula:

$$F1 = 2 * \frac{P * R}{P + R} \tag{4}$$

The MA F1 score calculates the F1 score separately for each class and then takes the average, treating both the classes equally while the WA F1 score calculates the F1 score for each class independently but when it averages them, it uses a weight equal to the number of true instances for each class. Similarly the MA precision and recall calculates takes the precision and recall values for both the classes and averages them out, treating them equally, while the WA precision and recall gives the average of both the classes but assigns a weight equal to the number of true instances for each class.

In Table 2 below, the classification results for LR is presented for each of the four features. The model is able to achieve an accuracy of 66% when MFCCs are used as features and an F1 score of 0.15 and 0.79 for Depressed (D) and

Non-Depressed (ND) class. It achieves a precision value of 0.12, 0.85 and a recall value of 0.21, 0.74 for depressed and non-depressed class respectively. The other three features are showing highly biased results. Even though they are showing a better accuracy compared to MFCCs, the F1 scores indicates their bias towards the non-depressed class. Since, LR assumes a linear boundary between the features and target class it may not be effective in capturing the complex decision boundaries, limiting its performance on complex relationships present in the AM and FM RFs. In Table 3, we have tabulated the classification results

Table 2: Classification results for LR

Features	**Accuracy**	**F1 Score**		**F1 Score**	**F1 Score**	**Precision**		**Precision**	**Precision**	**Recall**		**Recall**	**Recall**
		D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**
MFCCs	66%	0.15	0.79	0.47	0.70	0.12	0.85	0.48	0.75	0.21	0.74	0.48	0.67
FM RFs	86%	0.06	0.93	0.49	0.80	1.0	0.86	0.93	0.88	0.03	1.0	0.52	0.86
AM RFs	85%	0.0	0.92	0.46	0.79	0.0	0.86	0.43	0.74	0.0	1.0	0.50	0.86
AM+FM RFs	81%	0.04	0.92	0.47	0.78	0.08	0.86	0.47	0.75	0.03	0.94	0.49	0.82

for random forest classifier for all the four feature vectors. The random forest achieves an accuracy of 85%, 83%, 80% and 83% for MFCCs, FM RFs, AM RFs and the combination of AM and FM RFs, respectively. Similar to LR, random forest also turns out to be biased towards the non-depressed class as can be observed from the F1 scores of both the classes which are 0.0 in case of MFCCs and the combination of AM and FM RFs and 0.09 and 0.08 for FM RFs and AM RFs, respectively for the depressed class. Random forest is capable of modeling non-linear relationship by combining multiple DTs, however it introduces complexity and hence, no substantial improvements in F1 score is achieved as can be noticed from the results in Table 3.

Table 3: Classification results for random forest

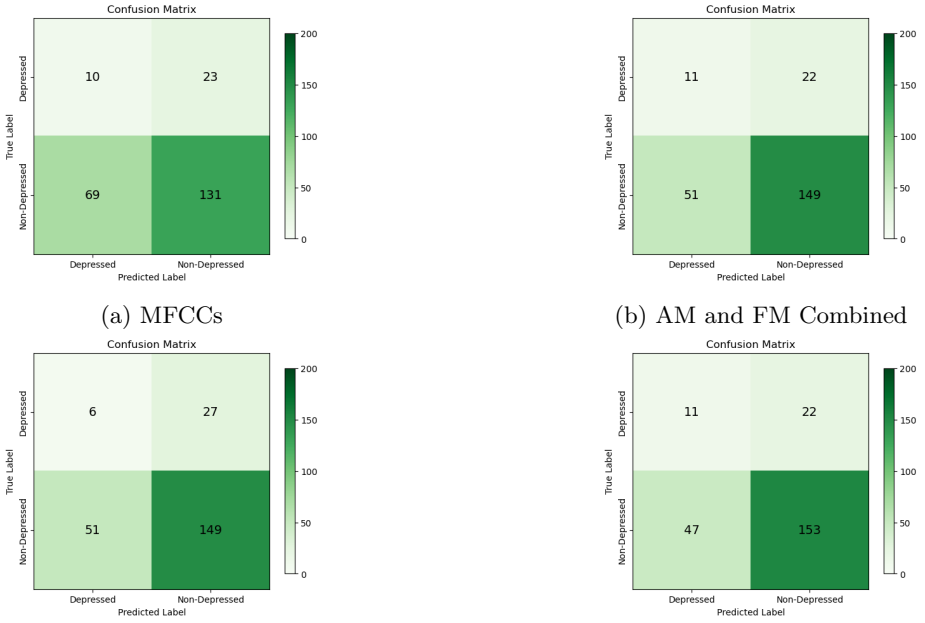
Features	**Accuracy**	**F1 Score**		**F1 Score**	**F1 Score**	**Precision**		**Precision**	**Precision**	**Recall**		**Recall**	**Recall**
		D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**
MFCCs	85%	0.0	0.92	0.46	0.79	0.0	0.86	0.43	0.74	0.0	0.99	0.50	0.85
FM RFs	83%	0.09	0.91	0.50	0.79	0.20	0.86	0.53	0.77	0.06	0.96	0.51	0.83
AM RFs	80%	0.08	0.89	0.48	0.77	0.12	0.86	0.49	0.75	0.06	0.93	0.49	0.80
AM+FM RFs	83%	0.0	0.91	0.45	0.78	0.0	0.85	0.43	0.73	0.0	0.97	0.48	0.83

Table 4 below shows the classification results for the DT model corresponding to all the features. The DT turns out to be most effective among all the classifiers used. The DT achieves an accuracy of 60%, 70%, 66%, 68% for MFCCs, FM RFs, AM RFs and the combination of AM and FM RFs, respectively. The model also achieves an improved F1 score for each of the features out of which the FM RF shows the most improved results with an F1 score of 0.24 and 0.82 for Depressed (D) and Non-Depressed (ND) classes, respectively. The confusion matrix shown

in Fig.8 (d) below suggests that in case of FM RFs, the DT classifier is able to predict the highest number of samples correctly compared to other features. Out of 200 non-depressed speech samples used for testing, it has correctly predicted 153 and out of 33 depressed speech samples, it is able to predicted 11 of them correctly. The confusion matrices for other feature are also presented in Fig.8 for the purpose of comparison. Since, the DT model is least biased and performing best with FM RFs, the classification system has been built using them.

Table 4: Classification results for DT

Features	**Accuracy**	**F1 Score**		**F1 Score**	**F1 Score**	**Precision**		**Precision**	**Precision**	**Recall**		**Recall**	**Recall**
		D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**	D	ND	**(MA)**	**(WA)**
MFCCs	60%	0.18	0.74	0.46	0.66	0.13	0.85	0.49	0.75	0.30	0.66	0.48	0.61
FM RFs	**70%**	**0.24**	**0.82**	**0.53**	**0.73**	**0.19**	**0.87**	**0.53**	**0.78**	**0.33**	**0.77**	**0.55**	**0.70**
AM RFs	66%	0.13	0.79	0.46	0.70	0.11	0.85	0.48	0.74	0.18	0.74	0.46	0.67
AM+FM RFs	68%	0.23	0.80	0.52	0.72	0.18	0.87	0.52	0.77	0.33	0.74	0.54	0.69



(a) MFCCs

(b) AM and FM Combined

(c) AM Only

(d) FM Only

Fig. 8: Confusion matrices showing the number of true and false prediction made by the decision tree classifier for (a) MFCCs, (b) A combination of AM and FM RFs, (c) AM RFs and (d) FM RFs for the test data

The improvement in the DT model can be explained based on its working principle. DTs have high decision boundary flexibility, allowing them to capture non-linear decision boundaries in the feature space in a much better way. By creating hierarchical decision rules, they are capable of handling the sequential data like AM and FM RFs which contains important time-frequency patterns.

6 Conclusions and Future Work

In this paper, AM and FM RFs and their combination have been studied for automatic classification of depression using speech data. Three different classifiers, i.e LR, random forest and DT were trained for each of the feature vectors and their performance was compared against MFCCs. FM RFs, when used as the feature vector for the DT classifier performed the best. By best, we mean that this system was least biased among the systems discussed in this paper in performing the classification task. In conclusion, RFs have the potential to capture the essential patterns from the speech data which may be indicative of depression. Experimentation on larger and balanced datasets could help establish its real-world applicability and reliability in aiding mental health assessments. Furthermore, using other relevant features along with the RFs may also have the potential to improve the system's performance.

References

1. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., et al.: From joyous to clinically depressed: Mood detection using spontaneous speech. In: FLAIRS Conference. vol. 19 (2012)
2. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech communication **71**, 10–49 (2015)
3. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE transactions on Biomedical Engineering **47**(7), 829–837 (2000)
4. Gibbon, D.: Speech rhythms: learning to discriminate speech styles. Proc. Speech Prosody 2022 pp. 302–306 (2022)
5. Gibbon, D.: The rhythms of rhythm. Journal of the International Phonetic Association **53**(1), 233–265 (2023)
6. Gibbon, D., Li, P.: Quantifying and correlating rhythm formants in speech. arXiv preprint arXiv:1909.05639 (2019)
7. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics **83**, 103–111 (2018)
8. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process **5**(2), 1 (2015)
9. Satt, A., Rozenberg, S., Hoory, R., et al.: Efficient emotion recognition from speech using deep learning on spectrograms. In: Interspeech. pp. 1089–1093 (2017)

10. Shen, Y., Yang, H., Lin, L.: Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6247–6251. IEEE (2022)
11. Yingthawornsuk, T., Keskinpala, H.K., Wilkes, D.M., Shiavi, R.G., Salomon, R.M.: Direct acoustic feature using iterative em algorithm and spectral energy for classifying suicidal speech. In: Eighth Annual Conference of the International Speech Communication Association (2007)
12. Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., Tao, J., Schuller, B.: Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. IEEE Journal of Selected Topics in Signal Processing **14**(2), 423–434 (2019)