Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 9701 – 9710

Research Article

Speech Emotion Recognition Using Ensemble Classifiers

M. Venkata Subbarao[1*], Sudheer Kumar Terlapu[2], G. Challaram[3], D. Girish Kumar[4]

Abstract

Human emotions play vital role in everyday communal relations. Speech Emotion Recognition (SER) has numerous applications such as clinical studies, commercial applications, entertainment, computer games and audio surveillance. For online learning/teaching emotional state of student's plays a crucial role. SER may help an instructor for developing the new strategies to manage student emotions in the classroom. Based on the motivations, the proposed work developed a SER system using Ensemble classifiers. Performance of the proposed ensemble techniques are tested under different noisy conditions and training rates. Thereafter, the accuracy of proposed techniques compared with traditional techniques.

Keywords: *Bagged trees, subspace KNN, MFCC, GTCC.*

Introduction

Emotions are more powerful indications which conveys the psychological state of an individual. Emotion recognition (ER) may help the doctors/ psychologists to understand the emotional state of a patent, which is a primary step in treatment. ER has many applications such as inter-face with robots, clinical studies, computer games, audio surveillance, cardboard systems, banking, entertainment, commercial applications, etc. There are several ways to recognize the emotion of a person which includes analysis of electroencephalogram (EEG) [1], Electrocardiography (ECG) signals [2], Electrodermal activity [3], Speech signals [4], facial expressions [5] and body language [6]. EEG, ECG and Electrodermal activity-based ER models requires a huge number of sensors for collection of the psychological signals. Recently ER using speech signal has gained the attention because of its advantages over other approaches. The computational complexity and hardware requirements in this approach is less than other approaches. The more common emotions are anger, joy, fear, disgust, sadness, boredom, and neutral.

[1*] Dept. of ECE, Shri Vishnu Engineering College for Women, India. E-mail: mandava.decs@gmail.com, ORCID 0000-0001-5840-2190

[2] Shri Vishnu Engineering College for Women, India.

[3] Shri Vishnu Engineering College for Women, India.

[4] Shri Vishnu Engineering College for Women, India.

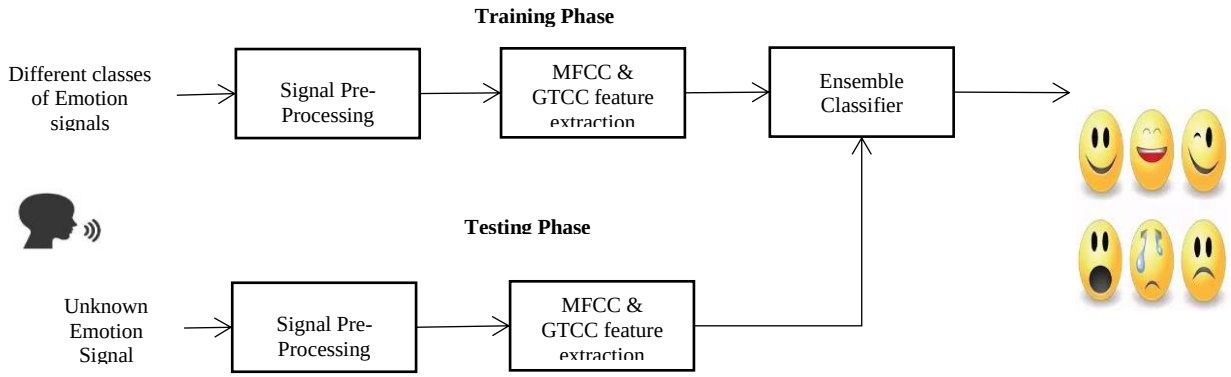


Fig. 1. Proposed SER System Framework

Recently, several methods have been developed to recognize the emotional state of a person from speech utterances. Existing approaches depend on extracting specific acoustic features from speech utterances such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), etc and then classifying the speech using traditional classifiers like Gaussian Mixture & Hidden Markov Models (GMM, HMM). The performance of the classifiers couldn't be improved beyond a certain limit [7]. Some of the recent works on SER is tabulated in Table I.

Table 1
SER Systems

Ref. No.	Features	Classifier	No. of Emotions	Accuracy (%)
8	MFCC	DNN- HMM	7	74.28/ 77.92
8	MFCC	GMM-HMM, SVM	7	76.18
9	MFCC	SNN	6	83
10	MFCC Acceleration, Velocity	DT CNN LSTM	7	80
11	Prosody, Quality	SVM	4	80.75
12	linguistic and non-linguistic	SVM, DT	6	84.39

From the literature, it is observed that the percentage of accuracy attained is about 80 and the performance is verified at only for fixed training rate. With these inputs, this paper analyses the accuracy with several training rates.

The organization of the paper is as follows. The framework & feature extraction is presented in Section 2. Methodology of the ensemble classifiers is presented in Section 3. Simulation results are discussed in Section 4. Conclusion of the work is discussed in Section 5.

Frame Work

The frame work of the proposed SER system is shown in Fig. 1. It involves Training and testing phases. In this work MFCC, GTCC features are considered to analyze the performance. The extraction of MFCC and GTCC features are shown in Fig. 2 and 3 respectively.

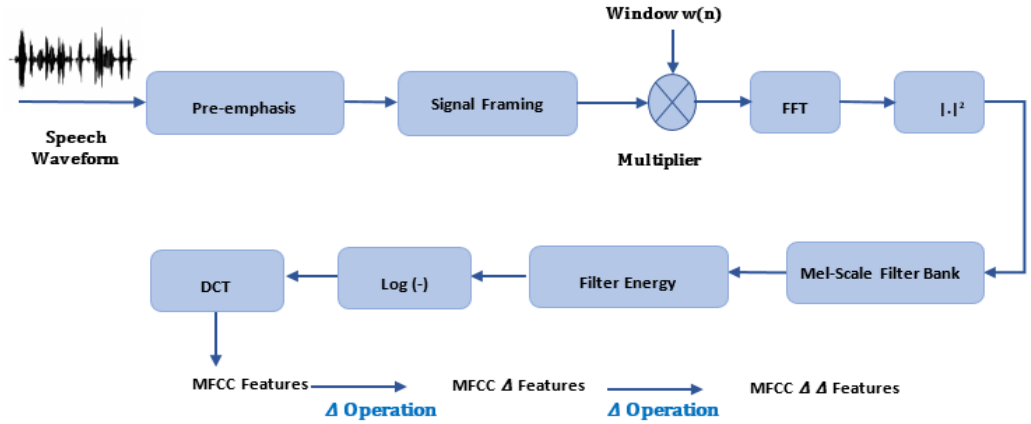


Fig. 2. MFCC Features Extraction

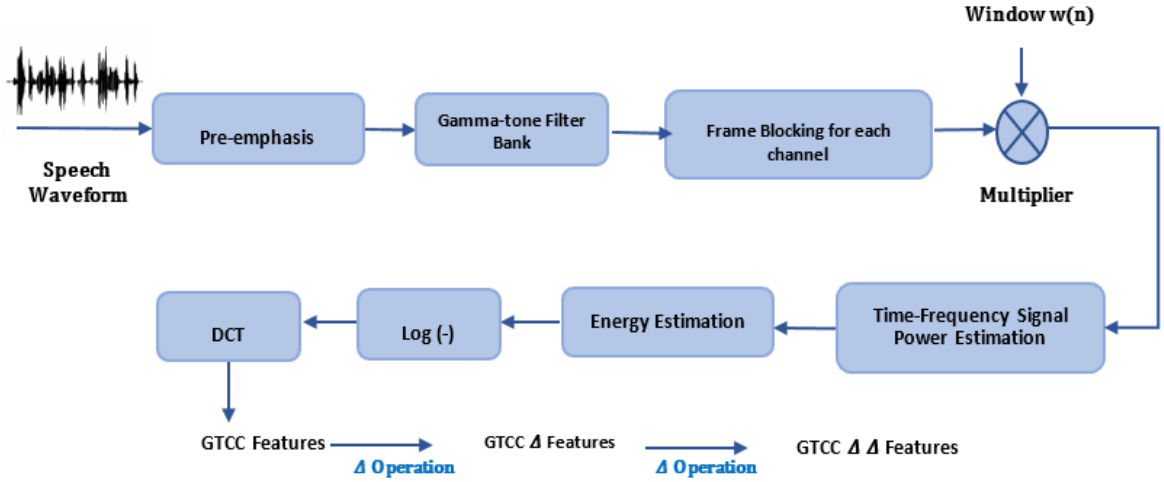


Fig. 3. Extraction of GTCC Features

Ensemble Classifiers

The pattern recognition classifiers such as SVM, DT, and KNN classifiers have different classification accuracy for the same data set. This is because of the algorithmic variation. Sometimes, it is very difficult to identify the best classifier among all for a given data set. The best remedy for this problem is construction of an ensemble classifier with a set of classifiers and take the majority vote for best prediction. Ensemble classification is a non-linear ML approach that selects a collection, or ensemble, of hypotheses from the hypothesis space and merges their predictions [13].

In this paper, bagged tree and subspace KNN based ensemble classifiers are developed for speaker emotion prediction. The block diagrammatic representation of proposed ensemble classifiers is shown in Fig. 4 [14]. Unlike boosting in bagging is a single step process and it reduces the variance by retaining the same forest for several times. In bagged trees each tree built autonomously. The emotion prediction is identified based on the weighted averages of the classification scores of the trees in bagging.

The weighted average of the selected trees is given by

$$\hat{P}_{avg}(y/x) = \frac{1}{\sum_{i=1}^{N} w_i z(i \in S)} \sum_{i=1}^{N} w_i \hat{P}_i(y/x)\, z(i \in S) \quad (1)$$

Where, N is the number of trees, $\hat{P}_i(y/x)$ is probability of class *y* given observation *x* using tree *i*. $z(i \in S)$ is 1 if *i* is in the set of *S*, else it is 0 and w_i is the weight of i[th] tree.
And the prediction class is given by the largest weighted average.

$$\hat{C}_{bag} = \underbrace{\arg max}_{y \in D} \{\hat{P}_{bag}(y/x)\} \quad (2)$$

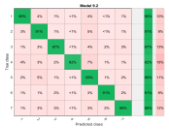
Where D is the set of all emotions.
The other side in subspace KNN classifiers use random subspace ensembles to improve the accuracy of KNN classifiers. These classifiers require very less memory than other classes of ensemble classifiers and also these can capable of handle the missing values (NaNs).

Simulation Results

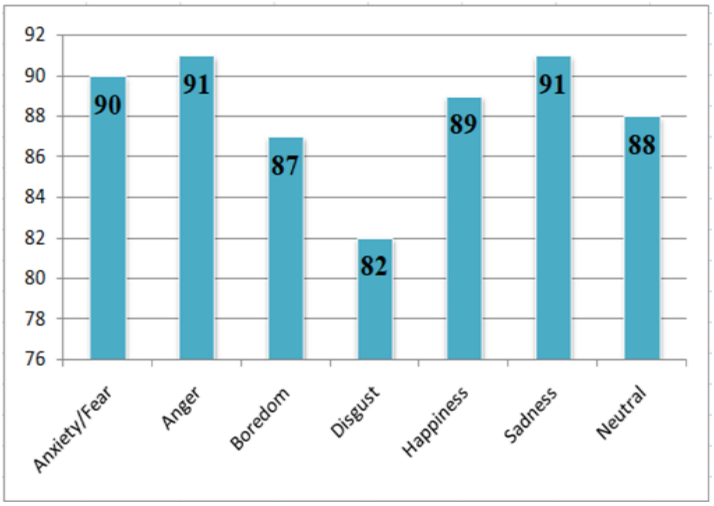
The simulation parameters for analysis of SER using bagged trees and subspace KNN classifiers is shown in Table II. For simulation berlin data set is considered and it consists of seven emotions (Anger, Anxiety/Fear, Boredom, Happiness, Disgust, Neutral and Sadness). To train the classifiers 39 MFCC and 39 GTCC features are extracted which includes Δ, ΔΔ features of both types. The performance is analyzed with training rates of 90-50%.

Table II
Simulation Information

Parameter	**Description**
Emotions	Anxiety, Anger, Boredom, Disgust, Happiness, Sadness, Neutral
Features	MFCC, MFCC Δ, MFCC ΔΔ, GTCC, GTCC Δ, GTCC ΔΔ
Size of Dataset	80000*78
SNR	0-40 dB
Training Rate (%)	90-50 with step size of 5



(a) Confusion Matrix



(b) Accuracy of each Emotion

Fig. 4. Performance of Bagged Trees at 90% Training Rate with MFCC Features

The performance of proposed classifiers is compared with some of the recent techniques is shown in Table III. From the simulation results it is clearly depicted that the proposed bagged tree & subspace KNN based classifiers are superior than existing approaches from the literature. The proposed approaches achieved the accuracy more than 90%.

Table III

Simulation Information

Ref. No.	Classifier	No. of Emotions	Accuracy (%)
9	SNN	6	83
10	DT CNN LSTM	7	80
11	SVM	4	80.75
12	SVM, DT	6	84.39
Proposed Ensemble Method	Bagged Trees Subspace KNN	7	92.5 (Max) 90.8 (Max)

Fig. 5 shows the performance of bagged tree ensemble classifier at 90% training rate. from the figure it is observed that among seven emotions disgust accuracy is below than 85%. The overall accuracy of bagged tree at 90% training rate is 88.2 % with MFCC and 90.8% with GTCC features. Similarly, the performance of bagged tree at various training rates is tabulated in Table IV.

Table IV

Performance Measure of Bagged Tree Classifier

Training Rate %	**True Class**	**Performance with MFCC, MFCC Δ, MFCC ΔΔ (39) features**								**Performance with GTCC, GTCC Δ, GTCC ΔΔ (39) features**							
		1	***2***	***3***	***4***	***5***	***6***	***7***	***Accuracy (%)***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Accuracy (%)***
90	1	90	4	1	<1	4	<1	1	88.2	84	3	2	8	<1	4		

	2	3	91	1	<1	5	<1	1		<1	91	2	5	<1	1	<1	
	3	1	3	87	<1	4	2	3			1	95	3	<1	1		
	4	4	3	2	82	7	1	1		1	2	1	94	<1	2	<1	90.8
	5	2	5	1	<1	89	1	2		2	3	4	5	83	3	<1	
	6	1	1	2	<1	3	91	2		<1	1	<1	2	<1	95	<1	
	7	1	3	3	<1	3	2	88		1	5	1	9	1	3	80	
80	1	86	6	1	<1	6	<1	2	87.0	81	4	2	12	1	4	1	88.2
	2	3	91	1		5	<1	2		1	88	3	6	<1	2	<1	
	3	1	3	84		5	2	4		<1	1	94	4	1	<1	<1	
	4	6	6	2	72	10	2	2		2	3	2	89	<1	4	<1	
	5	3	5	1	<1	88	1	2		3	2	9	9	79	4	<1	
	6	<1	1	3	<1	2	90	3		1	2	1	7	<1	89	<1	
	7	1	3	3		5	2	86		1	7	1	12	1	7	76	
70	1	85	5	1	<1	7	<1	1	83.8	76	4	2	12	1	4	1	85.5
	2	3	88	1	<1	6	<1	2		1	88	3	6	<1	2	<1	
	3	2	4	79	<1	6	3	7		<1	1	93	4	1	<1	<1	
	4	7	6	3	68	10	2	4		2	3	2	89	<1	4	<1	
	5	4	6	2		85	1	2		3	2	9	9	74	4	<1	
	6	1	3	4		3	85	5		1	2	1	7	<1	89	<1	
	7	2	4	4	<1	6	3	82		1	7	1	12	1	7	71	

60	1	82	6	1	<1	8	<1	2	79.8	72	5	3	14	1	4	1	82.6
	2	4	84	1	<1	8	<1	2		1	86	3	7	<1	2	<1	
	3	2	5	76	<1	6	3	8		<1	2	91	5	1	1		
	4	10	10	3	59	13	2	3		2	4	3	87	<1	4	<1	
	5	4	7	2	<1	83	1	3		3	3	11	10	69	5	1	
	6	1	3	4		5	80	6		1	2	1	9	<1	87	<1	
	7	2	4	6	<1	7	3	78		3	7	1	17	2	7	64	
50	1	76	9	2	<1	11	1	2	75.9	67	5	4	17	2	5	<1	79.5
	2	6	79	2	<1	11	<1	2		2	84	3	8	<1	3	<1	
	3	3	5	73	<1	8	3	8		1	2	89	6	1	1	<1	
	4	11	10	5	47	16	3	8		2	5	3	84	1	4	<1	
	5	5	9	3	<1	79	2	3		4	3	10	11	67	4	<1	
	6	1	3	5		5	79	7		1	3	1	10	1	84	<1	
	7	2	5	7	<1	8	3	74		3	10	1	20	2	7	57	

The performance of subspace KNN classifiers at various training rates is tabulated in Table V. From this it is observed that even at 50% training rate the performance is more than 78%. The performance of the proposed classifier with GTCC features is superior than that of MFCC features.

Table V
Performance Measure of Subspace KNN Classifier

Training Rate %	**True Class**	**Performance with MFCC, MFCC Δ, MFCC ΔΔ (39) features**								**Performance with GTCC, GTCC Δ, GTCC ΔΔ (39) features**							
		1	***2***	***3***	***4***	***5***	***6***	***7***	***Accuracy (%)***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Accuracy (%)***

90	1	91	3	1	<1	3	1	1	90.1	89	2	2	4	1	2	<1	92.5	
	2	2	92	1	<1	4	<1	1		1	93	2	3	<1	1	<1		
	3	1	3	88	<1	4	1	3		<1	1	94	2	1	1			
	4	4	4	2	84	4	<1	2		<1	2	1	94	1	2	<1		
	5	1	4	1	<1	91	1	2		1	2	3	4	88	1	1		
	6	1	1	1	<1	3	92	2		<1	1	<1	2	<1	96	<1		
	7	1	2	3	<1	4	1	89		<1	6	1	5	<1	3	85		
80	1	88	4	1	<1	5	<1	1	88	91	4	1	<1	5	<1	1	89.9	
	2	2	90	1	<1	4	<1	1		1	90	1	<1	4	<1	1		
	3	1	4	86	<1	5	1	3		1	4	89	<1	5	1	3		
	4	4	6	2	79	7	1	2		2	6	1	79	7	1	1		
	5	2	4	2	<1	88	1	2		2	4	1	<1	88	<1	1		
	6	1	2	2	<1	3	90	1		1	2	1	<1	3	91	1		
	7	1	3	2	<1	4	1	88		1	3	2	<1	4	1	90		
70	1	85	5	1	<1	6	1	1	85	84	3	3	7	1	2	1	89.3	
	2	2	87	1	<1	6	1	2		1	92	1	3	<1	2	<1		
	3	2	5	82	<1	5	2	4		1	2	92	3	1	1	<1		
	4	6	7	3	72	8	1	3		1	3	1	91	1	3	<1		
	5	2	5	2	<1	87	1	3		1	2	5	6	83	3	1		
	6	1	2	3	<1	4	86	3		1	2	1	5	<1	91	<1		
	7	2	4	4	<1	6	2	81		2	4	1	8	1	3	81		
60	1	82	6	2	1	8	1	2	80.9	81	3	2	9	1	3	1		
	2	3	83	2	1	8	1	3		1	91	1	5	<1	2	<1		

	3	3	5	78	<1	7	2	5		1	2	91	5	1	1	<1	
	4	6	10	4	64	12	1	3		2	4	2	88	1	3	1	86.5
	5	3	6	2	<1	84	1	3		2	3	5	7	80	3	1	
	6	2	3	4	<1	5	82	4		<1	3	1	6	1	89	<1	
	7	2	5	4	<1	7	2	78		2	8	1	9	1	4	74	
50	1	79	7	2	<1	9	1	2	77.8	76	5	3	10	2	3	1	84.4
	2	4	80	2	<1	10	1	3		1	89	2	5	1	2	1	
	3	2	6	75	<1	7	2	6		1	2	89	5	2	2	<1	
	4	7	12	5	56	13	1	6		2	4	2	87	1	3	1	
	5	4	7	3	<1	81	2	3		2	3	5	8	77	3	1	
	6	1	3	5	1	5	81	4		1	3	1	8	1	86	<1	
	7	3	6	5	<1	8	3	75		2	10	2	12	1	5	70	

Conclusion

Emotion recognition with two new ensemble classifiers was proposed. The investigations are carried out with two different sets of features i.e., MFCC, GTCC and their derivatives. From the simulation results it observed that GTCC features are superior than the MFCC features for SER. Thereafter, investigations are carried with the different training rates to know the effect of training rate on the performance. The proposed bagged tree and subspace KNN classifiers produced more than 75% accuracy even at 50% training rate. subspace KNN ensemble classifiers performed better than bagged tree ensemble classifiers in SER. Most of the existing approaches achieved 80% accuracy at 90% training rate but the proposed ones achieved more than 90% accuracy.

References

1. S.M. Alarcão and M. J. Fonseca.: Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing. 10* (3) (2019) 374-393.
2. Atefeh Goshvarpour, Ataollah Abbasi, Ateke Goshvarpour: An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical Journal. 40* (6) (2017) 355-368.
3. Ganapathy, N., Veeranki, Y.R., Kumar, H. et al. Emotion Recognition Using Electrodermal Activity Signals and Multiscale Deep Convolutional Neural Network. *J Med Syst*., 45, 49 (2021).

4. Kasiprasad Mannepalli, Panyam Narahari Sastry, Maloji Suman: Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University - Computer and Information Sciences.* (2018).
    a. doi.org/10.1016/j.jksuci.2018.11.012.
5. Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz J. Rak: Emotion recognition using facial expressions. *Procedia Computer Science.* 108 (2017) 1175-1184.
6. Felice Loi, Jatin G. Vaidya, Sergio Paradiso: Recognition of emotion from body language among patients with unipolar depression. *Psychiatry Research.* 209 (1) (2013) 40-49.
7. S. Zhang, S. Zhang, T. Huang, and W. Gao: Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia, 20*(6) (2018) 1576–1590.
8. L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli: Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition. *In Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.* (2013) 312–317.
9. E. Mansouri-Benssassi and J. Ye: Speech Emotion Recognition with Early Visual Cross-modal Enhancement Using Spiking Neural Networks. *In Proc. of 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary. (2019)* 1-8.
    a. doi: 10.1109/IJCNN.2019.8852473.
10. S. Basu, J. Chakraborty and M. Aftabuddin: Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. *In Proc. of 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore.* (2017) 333-336. doi: 10.1109/CESYS.2017.8321292.
11. Z. Han and J. Wang: Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine. *In Proc. of 2017 Chinese Automation Congress (CAC), Jinan. (2017)* 2513-2516. doi: 10.1109/CAC.2017.8243198.
12. J. Mao, Y. He and Z. Liu: Speech Emotion Recognition Based on Linear Discriminant Analysis and Support Vector Machine Decision Tree. *In proc. of 2018 37th Chinese Control Conference (CCC), Wuhan.* (2018) 5529-5533. doi: 10.23919/ChiCC.2018.8482931.
13. Subbarao M.V., Terlapu S.K., Chakravarthy V.V.S.S.S., Satapaty S.C.: In: Chowdary P., Chakravarthy V., Anguera J., Satapathy S., Bhateja V. (eds) Pattern Recognition of Time-Varying Signals Using Ensemble Classifiers. Microelectronics, Electromagnetics and Telecommunications. *Lecture Notes in Electrical Engineering,* 655 ((2021) 725-733. https://doi.org/10.1007/978-981-15-3828-5_76.
14. Venkata Subbarao, M., Samundiswary, P.: Performance Analysis of Modulation Recognition in Multipath Fading Channels using Pattern Recognition Classifiers. *Wireless Pers Commun.* 115 (2020) 129–151. https://doi.org/10.1007/s11277-020-07564-z.