Audio-visual Human Emotion Recognition Using Hierarchical Approach

Imran Khan[a], Sana Ul Haq[a], Muhammad Imran Majid[b]*, Imtiaz Rasool[a] and Muhammad Saeed Shah[a]

[a]Department of Electronics, University of Peshawar, Peshawar-25120, Pakistan

[b]Department of Electrical Engineering, Institute of Business Management, Karachi-75190, Pakistan

(received December 8, 2023; revised January 4, 2024; accepted January 4, 2024)

Abstract. This paper presents automatic human emotion recognition from audio-visual data. Both the hierarchical and flat approaches were implemented to obtain higher classification performance. The hierarchical approach was based on Mahalanobis distance. The Interactive Emotional dyadic Motion Capture database (IEMOCAP) was acquired and six different emotions, i.e., anger, excited, frustration, sadness, happiness and neutral state were used for the analysis. The method consisted of feature extraction, normalization, different feature selection and classification techniques. For flat approach, the best accuracy of 95.60% was obtained with Support Vector Machine (SVM) classifier and Info Gain feature selection. In the case of hierarchical approach, the best accuracy of 97.53% was achieved with Random Forest classifier and Correlation-based Feature Selection (CFS).

Keywords: audio-visual emotion recognition, feature extraction, hierarchical classification, mahalanobis distance, human-computer interaction

Introduction

Emotions play a crucial role in nonverbal humans' interaction. Humans can easily perceive emotions, but it is very challenging for machines to realize and respond to human feelings. The development of emotion recognition systems will provide more natural and efficient communication between people and machines. It will empower the machines to behave like humans (Thushara and Veni, 2016). Investigators from various disciplines have contributed to automatic emotion recognition (Zeng *et al.*, 2009). It has several applications in various fields including health care, distance learning, security, robotics and entertainment (Pablo *et al.*, 2014).

Speech and facial expressions are the prime signals used to recognize human affective states (Zeng *et al.*, 2007). Facial expressions play a vital role in visual emotion recognition by contributing about 55% of emotions in humans' communication (Mehrabian, 1968). Earlier studies have mainly focused on recognizing emotions from single modality (Schuller *et al.*, 2003; Tian *et al.*, 2001). The overall performance of unimodal approaches has been observed to be lower as compared to multimodal techniques (Guan *et al.*, 2009).

The first step in affect detection is to have a good quality data. Databases in various modalities and different languages have been recorded for this purpose. Audio databases include TESS (Dupuis and Pichora-Fuller, 2011), berlin database of emotional speech (EMO-DB) (Burkhardt *et al.*, 2005) and AIBO database (Batliner *et al.*, 2004). Visual datasets incorporate the Cohn-Kanade (Kanade *et al.*, 2000) and FABO database (Gunes and Piccardi, 2006). Audio-visual databases include the GEMEP (Bänziger *et al.*, 2006), SAVEE (Haq *et al.*, 2009) and RAVDESS (Livingstone and Russo, 2018).

The acoustic features extracted for emotion recognition include energy, pitch, formants, speech rate, mel frequency cepstral coefficients (MFCCs) and linear prediction coefficients (LPCs). The examples of visual features are facial action units, head pose, facial markers, and Gabor wavelets. The appearance and geometric are the two kinds of visual features associated to facial expressions (Zeng *et al.*, 2009). Bartlett *et al.* (2006) presented a technique based on appearance features, while Pantic and Bartlett (2007) used the geometric features.

To remove the noise and unwanted data, feature selection and reduction methods are normally used. These methods reduce computational complexity, whereas boost the categorization performance of a recognition system. The common feature selection techniques include best-first (Gunes and Piccardi, 2005), sequential forward

*Author for correspondence;
E-mail: imran.majid@iobm.edu.pk

selection (Haq *et al*., 2008) and greedy stepwise (Ranjan *et al*., 2021). The feature reduction techniques incorporate principal component analysis (PCA) (Fan *et al*., 2013) and linear discriminant analysis (LDA) (Haq *et al*., 2008).

The various emotion recognition schemes can be broadly classified in two sets: flat and hierarchical. In flat approach, the same set of features is used to classify all emotions at once. The disadvantage of using the same set of features is that it may effectively separate some emotions, but may not be able to distinguish between closely related emotion classes. In the case of hierarchical approach, emotion classes are separated stepwise. In the first step, all emotions are divided in two groups. In the next step, each binary class is further divided into other two classes. The procedure persists up to the separation of all emotions. At each step, different set of features is used to separate the binary classes. The hierarchical approach can effectively classify the more confusing classes because of using a different set of features at each binary level of classification. Researchers have used various classifiers for emotion classification including SVM (Lin and Wei, 2005), long short-term memory (LSTM) network (Araño *et al*., 2021), support vector neural network (SVNN) (Mannepalli *et al*., 2022) and convolutional neural network (CNN) (Alluhaidan *et al*., 2023).

The ensemble classifier technique has been suggested by some researchers to achieve better classification performance. Mohan *et al.* (2023) combined the 2D CNN and eXtreme grading boosting (XG-Boost) to accomplish an accuracy of 96.5% on the RAVDESS database using 16 emotion classes. The MFCC features were used for the classification. Bhanusree *et al.* (2023) utilized the CNN for feature extraction and random forest for classification. The proposed method achieved a recognition accuracy of 90.3% on the IEMOCAP and 92.2% on the RAVDESS databases. A hybrid model recommended by Badr *et al.* (2021) was consisted of convolutional and LSTM (ConvLSTM) networks. The proposed model obtained an accuracy of 91.0% on the RAVDESS dataset. Novais *et al.* (2022) used adaptive boosting (AdaBoost), neural network and random forest for speech emotion recognition. A majority vote based ensemble method was also explored. The random forest classifier obtained a recognition accuracy of 75.6% on the RAVDESS dataset. The individual classifiers performed better in comparison to the ensemble technique. Chalapathi *et al.* (2022) used acoustic features with AdaBoost classifier for speech emotion recognition. A classification score of 94.8% was obtained for seven classes on the RAVDESS database. Er (2020) used the acoustic and deep features with SVM classifier to recognize emotions from speech. The classification accuracies of 79.4%, 85.4% and 90.2% were achieved for the RAVDESS, IEMOCAP, and EMO-DB datasets, respectively.

This research aims to investigate the advantage of hierarchical approach over the flat approach. In addition, a bimodal approach was adopted to achieve better classification accuracy. The following sections present the IEMOCAP database, methodology, experimental results and discussion and conclusion.

IEMOCAP Database. IEMOCAP is an audio-visual emotional database recorded at the University of Southern California (Busso *et al*., 2008). The data was recorded from 5 males and 5 females. Each session involved a male and a female. A total of 53 markers were placed on the face, 2 markers on the headband, 2 markers on each wristband and an extra marker on each hand was also included.

The database has about 12 h of recordings. It contains 2066 improvised and 1761 scripted sentences, which sums to 3827. The database comprises both the scripted (5255 turns) and spontaneous (4784 turns) sessions. The actors recorded both the selected and improvised scripts in 10 emotions, i.e., anger, fear, disgust, frustration, happiness, excited, surprise, sadness, neutral state and other. The data was evaluated by 3 subjects and labeled based on majority vote.

The distinguished attributes of IEMOCAP database are its sufficient size, detailed capture information and true emotions elicitation method. In this research, IEMOCAP database is used for the analysis.

Material and Methods

The bimodal emotion recognition was comprised of the following steps: feature extraction, normalization, feature selection and classification. Both the flat and hierarchical approaches were used for the classification of emotions.

Feature extraction. In feature extraction the original raw data, e.g., audio, visual, is transformed into features. Audio features are obtained from speech signals, while visual features correspond to facial expressions and body language. The extracted audio features were related to Mel spectrum, signal energy, cepstral, spectral, raw signal, pitch and voice quality. The facial features related to roundness, angles, length and width of different parts of the face were extracted from facial marker points. The openSMILE (Eyben *et al.*, 2009) and MATLAB (Ljung, 2013) were used to extract a total of 7033 audio-visual attributes including 6539 audio and 494 visual features.

Feature normalization. The extracted features normally have different ranges of values because they are of different types. For this reason, feature normalization to a uniform range is essential for equal weighting of the various types of attributes. Feature normalization can be performed using the Weka toolkit (Witten *et al.*, 2010). The Z-Score and Min-Max normalization (Pandey and Jain, 2017) are the examples of feature normalization methods. In this research, Min-Max normalization with range [0, 1] was used.

The Min-Max normalization in the range [r_{min} r_{max}] is defined by the following equation

$$\overline{k} = \frac{k - k_{min}}{k_{max} - k_{min}} \times (r_{max} - r_{min}) + r_{min} \text{ ....(1)}$$

where:

$\overline{k}$, k_{min} and k_{max} are the normalized, minimum and maximum values of attribute *k*.

Feature selection. Feature selection is required to eliminate the redundant and unrelated features from the extracted set of features. Features can be selected either as a subset or individual features can be ranked based on some criterion. In this research, both the feature subset and individual feature ranking methods were implemented using Weka software. The feature subsets were selected using CFS evaluator with Best First and Greedy Stepwise search methods. Whereas the individual attributes were rated using the Info Gain and Gain Ratio attribute evaluators. The Weka toolkit was used for feature selection.

Classification. The bimodal emotion recognition was accomplished using 5 classifiers, i.e., Bayes Net, SVM, bagging, random forest and random tree. These classifiers utilize various methodologies for the categorization. The Bayesian classifiers are based on the Bayes' theorem. The probabilities of different classes are computed based on given features. SVM transforms the data to high dimensional space for transparent separation. It is faster and works well for high dimensionality and small training data. Bagging is an ensemble technique aspired to enhance the stability and accuracy of classification algorithms. It lowers the overfitting by reducing the variance. A random forest fits numerous decision trees on several sub-samples of the dataset. The classification accuracy is enhanced by employing the averaging. Random tree is an ensemble method of machine learning. The ensemble technique utilizes several base models to obtain the final prediction.

The classification experiments were conducted using both the flat and hierarchical approaches. A total of 6 emotion classes, i.e., anger, excited, frustration, sadness, happiness and neutral state, that had enough data were selected from the IEMOCAP for the experiments. The selected data contained 2778 sentences. The experiments were performed with 10-fold cross validation method using Weka software.

Flat approach. In flat approach, different emotion classes are separated using a single set of selected features. This technique may result in lower classification performance for the more confused emotion classes, as they are hard to separate while using same set of selected features for all emotions.

Hierarchical approach. In this technique a branched tree is constructed through binary classification. The hierarchical technique is believed to perform better than the traditional flat approach as a different set of features is used at each level of binary classification.

In this research, the hierarchical approach was used for bimodal emotion recognition based on Mahalanobis distance (Fan *et al.*, 2013). The Mahalanobis distance for two classes is given by the relation

$$d_{Mah} = \sqrt{(\mu_i - \mu_j)^T (P_i \Sigma_i + P_j \Sigma_j)^{-1} (\mu_i - \mu_j)} \text{ ....(2)}$$

where:

μ_i and μ_j represent the means, Σ_i and Σ_j represent the covariances, while P_i and P_j denote the prior probabilities of two normal distributed classes.

The Mahalanobis distance between each pair of emotions was computed. In this research, a total of six emotions, i.e., anger, excited, frustration, sadness, happiness, and neutral state, from the IEMOCAP database were used. Firstly, all these emotions were combined into a master class as shown in Fig. 1. In the next step, master class was split in class A and class B. Similar emotions were grouped together. Three emotions, i.e., anger, excited and frustration were placed in class A, while sadness, happiness and neutral state were placed in class B. Afterwards, class A was divided into subclasses A_1 and A_2. The class A_1 contains anger emotion, while class A_2 contains excited and frustration. Similarly, class B was separated into subclasses B_1 and B_2. In the last phase, subclasses A_2 and B_2 were further split into binary classes. All emotion classes were classified using this procedure.

Results and Discussion

The experiments were conducted using the audio-visual data of six emotions from the IEMOCAP database. The audio and visual features were merged at feature level. The results were averaged over 10-fold cross validation. Experiments were performed using both the flat and hierarchical classification approaches.

Flat approach. The classification results for the flat approach using different attribute evaluators and search methods are given in Table 1. In the case of CFS attribute evaluator with best first search method, the best accuracy of 92.76% was accomplished with Bayes Net classifier utilizing 86 features. The bagging and random forest classifiers also performed better, while SVM performed poorly. In the case CFS attribute evaluator with Greedy Stepwise search method, the recognition accuracy of 93.19% was attained with Bayes net classifier utilizing 110 features. Bagging and random forest classifiers also performed better, while the Random Tree performance was the lowest. The overall performance of CFS attribute evaluator for both the best first and Greedy stepwise search methods were quite close.

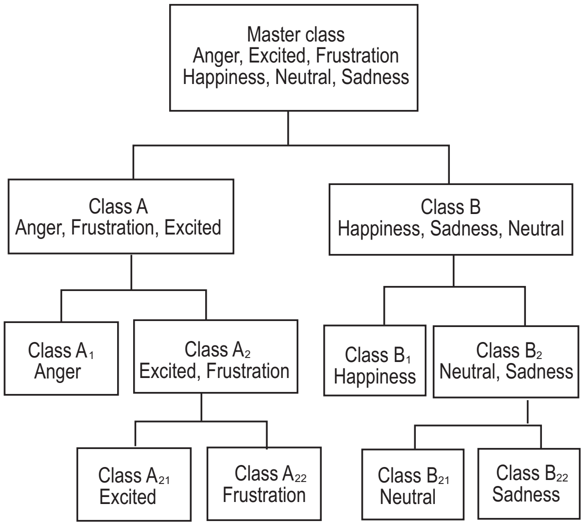

Fig. 1. Hierarchical tree based on Mahalanobis distance.

For the info gain and gain ratio attribute evaluators with Ranker search method, the classification tests were conducted for the top ranked features starting from 50 with a step size of 50. The best results for different classifiers were obtained using various numbers of attributes. In the case of Info Gain, the best recognition score of 95.60% was attained with SVM classifier using 2500 features. Other classifiers also performed well except the Random Tree. In the case of Gain Ratio, the top performance of 93.70% was accomplished with SVM classifier utilizing 3000 features. The Bagging classifier also performed better but other classifiers performed poorly.

For both info gain and gain ratio, the Bayes net, random forest and random tree classifiers provided higher accuracy for less number of attributes, while the SVM and Bagging accomplished better scores for larger number of attributes. The recognition performance of SVM and Bagging classifiers enhanced with growing the number of attributes, while that of other classifiers declined.

Hierarchical approach. The classification results for the hierarchical approach are given in Table 2. For the CFS attribute evaluator with the best first and greedy stepwise search methods, the random forest classifier provided the best results. The best classification result of 97.44% was obtained for the CFS with best first search method, while the best result of 97.53% was achieved for the CFS with Greedy stepwise search method. For both search methods, other classifiers also performed better but the SVM performance was poor.

For the Info gain and gain ratio attribute evaluators and ranker search method, the best accuracies of 96.52% and 95.94% were obtained, respectively using bagging classifier. Other classifiers also performed better for these attribute evaluators.

Comparison of flat and hierarchical approaches. The comparison of best classification results for the flat and hierarchical approaches are given in Table 3. For the CFS attribute evaluator with best first and Greedy stepwise search methods, the bayes net classifier provided the best results for the flat approach, while random forest performed better in the case of hierarchical approach. For the info gain and gain ratio attribute evaluators and ranker search method, the SVM classifier delivered the best outcomes for the flat approach, while bagging accomplished better scores in the case of hierarchical approach.

Table 1. Average classification accuracies (%) for the flat approach using different feature selection and classification techniques

Attribute evaluator	Search method	Number of	Classifiers				
		selected attributes	Bayes net	SVM	Bagging	Random forest	Random tree
CFS subset	Best first	86	92.76	79.22	91.28	91.43	82.57
	Greedy stepwise	110	93.19	82.82	91.54	91.03	79.98
Info gain	Ranker	2500	93.16 (200)	95.60	93.12 (1500)	94.63 (200)	86.28 (200)
Gain ratio	Ranker	3000	75.19 (150)	93.70	92.58	79.19 (150)	69.04 (150)

Table 2. Average classification accuracies (%) for the hierarchical approach using different feature selection and classification techniques

Attribute	Search	Classifiers				
evaluator	method	Bayes net	SVM	Bagging	Random forest	Random tree
CFS subset	Best first	95.63	84.14	95.29	97.44	94.79
	Greedy stepwise	95.50	84.23	95.27	97.53	94.71
Info gain	Ranker	91.19	92.36	96.52	95.72	91.84
Gain ratio	Ranker	87.11	92.47	95.94	94.61	90.31

Table 3. Comparison of best classification results for the flat and hierarchical methods

Attribute	Search	Flat method		Hierarchical method	
evaluator	method	Classifier	Accuracy (%)	Classifier	Accuracy (%)
CFS	Best first	Bayes net	92.76	Random forest	97.44
	Greedy stepwise	Bayes net	93.19	Random forest	97.53
Info gain	Ranker	SMO	95.60	Bagging	96.52
Gain ratio	Ranker	SMO	93.70	Bagging	95.94

The overall recognition accuracy of hierarchical approach was superior to flat approach. In the case of flat approach, the best recognition score of 95.60% was attained with SVM using Info gain attribute evaluator with ranker search method. For the hierarchical approach, the best result of 97.53% was acquired with random forest using CFS attribute evaluator and Greedy stepwise search method.

Conclusions

In this research, the bimodal emotion recognition was accomplished using both the flat and hierarchical approaches. The experiments were performed using six emotion classes of the IEMOCAP database. The audio and visual attributes were extracted. Feature selection was employed using the CFS, info gain and gain ratio attribute evaluators with best first, Greedy stepwise and ranker search methods. The emotions were classified using five types of classifiers, i.e., Bayes net, SVM, bagging, random forest and random tree.

For flat approach, the best recognition accuracy of 95.60% was attained with SVM classifier using Info gain attribute evaluator and ranker search method. For hierarchical approach, the best result of 97.53% was obtained with random forest classifier using CFS attribute evaluator and Greedy stepwise search method. The hierarchical approach accomplished better performance in comparison to flat approach.

In future, it will be exciting to explore the other distance measures, e.g., KL divergence measure and Bhattacharyya distance, for building the hierarchical tree. In addition, these research findings need to be validated using other audio-visual databases such as SAVEE.

Conflict of Interest. The authors declare that they have no conflict of interest.

References

Alluhaidan, A.S., Saidani, O., Jahangir, R., Nauman, M.A., Neffati, O.S. 2023. Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, **13:** 4750.

Araño, K.A., Gloor, P., Orsenigo, C., Vercellis, C. 2021.

When old meets new: emotion recognition from speech signals. *Cognitive Computation*, **13:** 771-783.

Badr, Y., Mukherjee, P., Thumati, S. 2021. Speech emotion recognition using MFCC and hybrid neural networks. In: *Proceedings of International Joint Conference on Computational Intelligence*, pp. 366-373, Valletta, Malta.

Bänziger, T., Pirker, H., Scherer, K. 2006. GEMEP-GEneva multimodal emotion portryals: a corpus for the study of a multimodal emotional expressions. In: *Proceeding of LREC Workshop on Corpora for Research on Emotion and Affect*, pp. 15-19, Genoa, Italy.

Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J. 2006. Fully automatic facial action recognition in spontaneous behaviour. In: *Proceeding of International Conference on Automatic Face and Gesture Recognition*, pp. 6-11, Southampton, UK.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M.J., Wong, M. 2004. You stupid tin box - children interacting with the Aibo robot: a cross-linguistic emotional speech corpus. In: *Proceedings of the International Conference of Language Resources and Evaluation*, pp. 171-174, Lisbon, Portugal.

Bhanusree, Y., Kumar, S.S., Rao, A.K. 2023. Time-distributed attention-layered convolution neural network with ensemble learning using random forest classifier for speech emotion recognition. *Journal of Information and Communication Technology*, **22:** 49-76.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B. 2005. A database of German emotional speech. In: *Proceeding of European Conference on Speech Communication and Technology*, pp. 3-6, Lisbon, Portugal.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **42:** 335-359.

Chalapathi, M.V., Kumar, M.R., Sharma, N., Shitharth, S. 2022. Ensemble learning by high-dimensional acoustic features for emotion recognition from speech audio signal. *Security and Communication Networks*, **2022:** 1-10.

Dupuis, K., Pichora-Fuller, M.K. 2011. Recognition of emotional speech for younger and older talkers: behavioural findings from the toronto emotional speech set. *Canadian Acoustics*, **39:** 182-183.

Er, M.B. 2020. A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, **8:** 221640-221653.

Eyben, F., Wöllmer, M., Schuller, B. 2009. openSMILE – the munich versatile and fast open-source audio feature extractor categories and subject descriptors. In: *Proceedings of International Conference on Affective Computing and Intelligent Interaction*, pp. 1-6, Amsterdam, Netherlands.

Fan, Z., Ni, M., Sheng, M., Wu, Z., Xu, B. 2013. Principal component analysis integrating mahalanobis distance for face recognition. In: *Proceedings of Second International Conference on Robot, Vision and Signal Processing*, pp. 89-92, Kitakyushu, Japan.

Guan, L., Muneesawang, P., Wang, Y., Zhang, R., Tie, Y., Bulzacki, A., Ibrahim, M.T. 2009. Multimedia multimodal methodologies. In: *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1600-1603, New York, USA.

Gunes, H., Piccardi, M. 2006. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: *Proceedings of International Conference on Pattern Recognition*, pp. 1148-1153, Hong Kong, China.

Gunes, H., Piccardi, M. 2005. Affect recognition from face and body: early fusion vs. late fusion. In: *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pp. 3437-3443, Waikoloa, USA. DOI: 10.1109/ICPR.2006.39

Haq, S., Jackson, P.J., Edge, J. 2008. Audio-visual feature selection and reduction for emotion classification. In: *Proceedings of International Conference on Auditory-Visual Speech Processing*, pp. 185-190, Moreton Island, Australia. DOI: 10. 1109/ICSMC.2005.1571679

Haq, S., Jackson, P.J. 2009. Speaker-dependent audio-visual emotion recognition. In: *Proceedings of International Conference on Audio-Visual Speech Processing*, pp. 1-6, Norwich, UK. https://openresearch.surrey.ac.uk/esploro/outputs/conferencePresentation/Speaker-dependent-audio-visual-emotion-recognition/99512249002346

Kanade, T., Cohn, J.F., Tian, Y. 2000. Comprehensive database for facial expression analysis. In: *Proceedings IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, Grenoble, France.

Lin, Y.L., Wei, G. 2005. Speech emotion recognition based on HMM and SVM. In: *Proceedings of International Conference on Machine Learning and Cybernetics*, pp. 18-21, Guangzhou, China.

Livingstone, S.R., Russo, F.A. 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:** e0196391.

Ljung, L. 2014. System Identification Toolbox: User's Guide, MathWorks Inc. https://www.mathworks.com/help/ident/

Mannepalli, K., Sastry, P.N., Suman, M. 2022. Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University-Computer and Information Sciences*, **34:** 384-397.

Mehrabian, A. 1968. Communication without words. *Psychology Today*, **2:** 53-56.

Mohan, M., Dhanalakshmi, P., Kumar , R.S. 2023. Speech emotion classification using ensemble models with MFCC. *Procedia Computer Science*, **218:** 1857-1868.

Novais, R.M., Cardoso, P.J., Rodrigues, J.M. 2022. Emotion classification from speech by an ensemble strategy. In: *Proceedings of International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, pp. 85-90, Lisbon, Portugal.

Pablo, J., Busso, C., Becerra, N. 2014. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech and Language*, **28:** 278-294.

Pandey, A., Jain, A. 2017. Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, **9:** 36-42.

Pantic, M., Bartlett, M.S. 2007. Machine analysis of facial expressions. In: *Face Recognition*, Delac, K., Grgic, M. (ed.), pp. 377-416, InTech Education and Publishing, Vienna, Austria. DOI: 10.5772/4847

Ranjan, A., Singh, V.P., Mishra, R.B., Thakur, A.K., Singh, A.K. 2021. Sentence polarity detection using stepwise greedy correlation based feature selection and random forests: an fRMI study. *Journal of Neurolinguistics*, **59:** 100985.

Schuller, B., Rigoll, G., Lang, M. 2003. Hidden Markov model-based speech emotion recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1-4, Hong Kong, China.

Thushara, S., Veni, S. 2016. A multimodal emotion recognition system from video. In: *Proceedings of International Conference on Circuit, Power and Computing Technologies*, pp. 1-5, Nagercoil, India. DOI: 10.1109/ICCPCT.2016.7530161

Tian, Y., Kanade, T., Cohn, J.F. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23:** 97-115.

Witten, I.H., Frank, E., Hall, M.A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*, 664 pp., Morgan Kaufmann, New York, USA. https://doi.org/10.1016/C2009-0-19715-5

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S. 2009. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31:** 39-58.

Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Roth, D., Levinson, S. 2007. Audio-visual affect recognition. *IEEE Transactions on Multimedia*, **9:** 424-428.