Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Acoustic Features Based Emotional Speech Signal Categorization by Advanced Linear Discriminator Analysis

Subarna Shakya

Professor, Department of Electronics and Computer Engineering, Central Campus, Institute of Engineering, Pulchowk, Tribhuvan University, Pulchowk, Lalitpur, Nepal

E-mail: drss@ioe.edu.np

Abstract

Personal computer-based data collection and analysis systems may now be more resilient due to the recent advances in digital signal processing technology. The signal processing approach known as Speaker Recognition, uses the specific information contained in voice waves to automatically identify the speaker. For a single source, this study examines systems that can recognize a wide range of emotional states in speech. Since it offers insight into human brain states, it's a hot issue in the development during the interface between human and computer arrangement for speech processing. Mostly, it is necessary to recognize the emotional state of people in the arrangement. This research analyses an effort to discern various emotional stages such as anger, joy, neutral, fear and sadness by classification methods. The acoustic feature, a measure of unpredictability, is used in conjunction with a non-linear signal quantification approach to identify emotions. The unpredictability of all the emotional signals is included in a feature vector constructed from the calculated entropy measurements. In the next step, the acoustic features through speech signal are used for the training in the proposed neural network that are given to linear discriminator analysis approach for further greater classification with acoustic feature extraction. Besides, this research article compares the proposed work with various modern classifiers such as K- nearest neighbor, support vector machine and linear discriminator approach. Moreover, this proposed algorithm is based on acoustic features in Linear Discriminant Analysis (LDA) with acoustic feature extraction machine algorithm. The great advantage of this proposed algorithm is that it separates negative and positive features of emotions and provides good results during classification. According to the results from efficient cross-validation in the proposed framework, accessible sample of dataset of Emotional

DOI: https://doi.org/10.36548/jscp.2021.4.005
Received: 20.11.2021, received in revised form: 19.12.2021, accepted: 27.12.2021, published: 07.01.2022

Speech, a single-source LDA classifier can recognize emotions in speech signals with above 90 percent of accuracy for various emotional stages.

Keywords: Speech signal, feature vector, Linear Discriminant Analysis, Machine learning algorithm, acoustic feature extraction

1. Introduction

A wide range of information is encoded in the voice signal. The word spoken is the primary means of delivering a message. Speech also conveys information about the speaker's emotions, gender, and identity. Speech recognition and automated speaker recognition go hand in hand. It is the purpose of automated speaker identification, rather than that of speech recognition, to recognize the speaker by extracting, characterizing, and recognizing the information included within the voice signal [1-5].

Speaker recognition technology has a wide range of uses and is constantly evolving. This technique comprises of speech signal volumes to verify the speaker’s identity at the current position. This control mechanism is used in many sectors such as voice dialing, phone banking etc. through various voice dataset in the well-developed countries. Besides, this application of techniques is used in security control in many defense sectors, forensic applications etc. New services based on speaker identification are on the way, and they promise to make our lives easier in the process.

Many authors have written extensively on the subject of computer-based analysis and automated assessment of disordered speech. This topic has shown more importance when it comes to pre-processing (non-standard) audio signals. Studying the extraction of certain features, which are used to categorize and evaluate patients, was the primary focus. Finally, there is an investigation into the substance of diagnostic and prognostic decision procedures for certain illnesses and treatments [6-10]. Speech signals have been processed, analyzed, classified, and recognized for a long time. Research on the keywords stated above may be readily accessed in the literature and includes both basic research and many application articles. It is possible to assess the sound quality of speech signals using a broad range of speech acoustics assessment techniques, which allow for a multidimensional analysis of the findings and features that reveal how they vary throughout articulation [11, 12]. It is, however, very difficult and time-consuming to analyze these aspects directly, particularly when it comes to pathological speech analysis. Despite this, it leads to improvements in techniques for process

control, analysis, and speech signal detection, and the outcomes of this study are documented in several publications [13–15].

1.1 Motivation of the Research

Many research have shown that emotional voices and acoustic characteristics are linked. Although there is no explicit and deterministic mapping between components and emotional state, spoken emotion recognition has a lower recognition rate than other emotion-recognition approaches. Therefore, it is crucial to discover the right mix of features for speech-based emotion detection.

2. Organization of the Research

Several sections comprise this research article as follows; section 2 provides a literature survey about speech signal categorization by various algorithms. Section 3 delivers the proposed work for classifying various emotional features through speech input signals. Section 4 discusses the obtained results of the proposed algorithm. Finally, the conclusion of speech analysis is addressed at the end of this research article.

3. Related Works

Binary decision trees were used by Wieman et al. to identify the traits most closely associated with emotions. However, just a tiny sample size was used for their study. It was considered that each person's speech qualities differed and that emotions were impacted by the speaker's age, gender, and acoustic characteristics. As a result, the researchers concentrated on emotion identification by aggregating speech data based on age and gender [16]. Gender and age-based hierarchical models were presented, using data drawn from OpenSmile [17] and eGeMAPS [18] feature vectors.

German utterances are documented in the Berlin Emotional Speech Database (EmoDB). In order to ensure that the database was as realistic as possible, it was built using simple phrases from the real world. For each of the seven types of emotions — rage, boredom, disgust, fear, happiness, sorrow, and neutral, 10 actors (five females and five males) recorded their speech data [19].

More than 12 hours of audio and video were captured using Emotional Dyadic Motion Capture (IEMOCAP). During each recording session, audio, face, and landmark data were

collected. Men and women conversed back and forth during each session. During the recording, 10 performers were divided into five pairs. To provide the highest possible sound quality, the whole project was recorded in a professional film studio. The actors were sitting three meters apart in a "social" position [20].

Emotional speech and song in North American English are categorized into eight different emotions: neutral, calm, joyful, sad, angry, afraid, disgusted, and startled. The database contains data from a wide range of professional performers, each of whom has a unique collection of audio-visual assets and musical compositions that may be used as data points. For each recorded performance by an actor, AV, video-only, and audio-only formats were available [21].

An automatic speech emotional recognition approach is based on a greater arrangement of characteristics gathered in signal processing domain was suggested by Kerkeni and colleagues. Based on the intermediate frequency signal, spectral and frequency properties mixed with cepstral components were employed [22].

4. Methodologies

4.1 K-Nearest Neighbour

When it comes to classification and prediction, K-NN is one of the most used algorithms. Due to the lack of specific training data, it is seen as a lazy learning strategy. In most cases, the complete dataset is used in training. Since there are no assumptions made while comparing characteristics to anticipate a new data point, it is also a non-parametric technique. The first selection of "K" neighbors might be any integer depending on the dataset's class count [23]. Distance metrics like as Euclidean, Hamming, etc. are used to determine the distance between training and test data. Class labels with the lowest calculated distance are shown first, and the information requested is ordered by distance.

4.2 Support Vector Machine

For classification and regression problems, Support Vector Machines (SVMs) are used [31]. For each data point, the feature value is represented as the coordinate value and plotted in a n-dimensional space. The two classes are separated by decision boundaries based on hyperplanes. Several options are explored in the search for the hyperplane, and finally the plane

with the greatest margin between the two classes is chosen. With absolute certainty, the separation plane identifies the test point for the future [24 - 27].

4.3 Proposed framework with Linear Discriminant Analysis

This proposed framework includes the standard workflow of speech pre-processing, which includes signals collection, pre-emphasis, and post-emphasis. The windowing with framing process is characterized for voice signal based on its acoustic characteristics. The adaptive signal decomposition is divided into positive and negative emotion characteristics, which are then used in the classification stage to achieve high levels of accuracy. Based on the acoustic features to LDA paired with gradient boosting machine process, attributes are retrieved and used in the LDA [28 – 31]. Figure 1 shows the overall proposed framework for emotional recognition from speech signals.

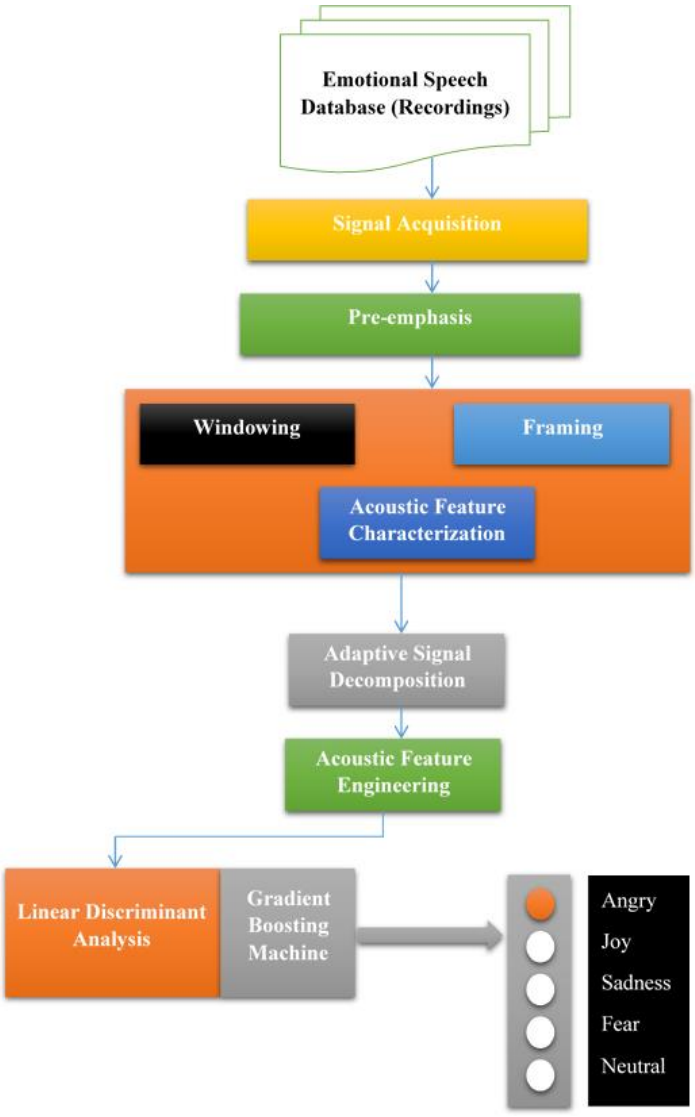

Figure 1. Overall proposed framework for emotional speech signal classification

4.3.1 Emotional categorization through adaptive signal decomposition

The machine learning method of linear discriminant analysis is well-known for classifying and forecasting problems. When compared to other classification algorithms, this method's prediction procedure is rather straightforward. LDA is a method for reducing dimensionality that works by reducing the size of higher-dimensional space. In LDA, the Intraclass variances defined as min, max and mean findings for the items of each period are used to assess the separability between the classes. Finally, Fisher's criteria reduce the variation within classes while increasing the distance between classes, resulting in a two-dimensional space. For Empirical Methods, the signals decomposed with adaptive manner into intermediate frequency approach is confined in the acoustic features. The information included in each intermediate frequency signal is not homogeneous and changes based on the input speech signal because various emotional states are recorded in unique features components of acoustics in speech signals. Examples of speech cues indicating good emotions include delight and pleasant surprise. Similarly, various frequency scales are used to capture negative emotions such as disgust and melancholy, as well as rage and terror [32]. Because of this, it is impossible to predetermine any intermediate frequency component or frequency scale.

4.3.2 Acoustic features extraction machine

The acoustics-based features are extracted and used to make prediction very effective. Each of these methods has its own strengths and weaknesses, but they all work well together to assist students who have difficulty making decisions to make the best choices possible. In each decision tree, the optimum split is determined by a separate set of attributes. As a result of this approach, each tree is distinct and capable of detecting additional signals in the underlying data. Another important aspect of the process is sequentially constructing new trees depending on the flaws of the preceding one.

5. Results and Discussion

One hundred and seventy audio signals of various acoustic characteristics are employed. Once all 229 recorded audio signals were decomposed, the 42-octave distribution values were recovered. LDA was then used to classify all 229 signals based on the feature sets retrieved from the extracted feature sets. Table 1 shows the various emotional recording speech inputs.

Table 1. Databases used for proposed algorithm

Emotions	Input Recordings (numbers)
Angry	30
Joy	25
Sadness	44
Fear	60
Neutral	70

Using the standard LDA with gradient boosting approach, all 229, 30 angry, 25 joy, 44 sadness, 60 fear and 70 neutral were accurately categorized with highest percent accuracy than other traditional approaches. However, two of the 2 fear and two of the 1 joy signals were incorrectly identified and tabulated in classification accuracy. Training by acoustic features that updates for emotional classification is tabulated in Table 2.

Table 2. Training by acoustic features update for emotion classification

S.No	Model	Min	Mean	Max
1	SVM	0.801	0.84	0.881
2	KNN	0.773	0.804	0.834
3	LDA	0.881	0.905	0.928
4	LDA + Boosting algorithm	0.931	0.951	0.991

The classification accuracy with the leave-one-out strategy was 91.23 percent with LDA alone, demonstrating the robustness of the proposed methodology and its independence from the dataset size. Many factors influence how well a series of frames capturing system works, including the kind of database it uses. The proposed Classification algorithms are evaluated on various kinds of speech signal databases that is showed in table 1. Speech signals for various emotions are created by skilled actors and stored in "acted databases," or simulated databases. Table 3 shows computed accuracy calculation results by the proposed algorithm.

Table 3. Obtained accuracy of classification results

Model	Emotion features based on acoustic phenomena				
	Angry	**Joy**	**Sadness**	**Fear**	**Neutral**
SVM	80.81%	79.23%	78.67%	82.62%	98%
KNN	82%	81.78%	83.68%	86.45%	98%
LDA	89.12%	89.89%	91.23%	87%	99%
LDA + Boosting algorithm	93.78%	95.12%	92.56%	94%	99%

These databases, known as elicited databases, are acquired from manufactured events that provoke new feelings of indifference. The misclassified signals were studied, but no audible clues were found as to why they were incorrectly classified as signals of unknown quality or origin. Table 4 shows the observation conclusion of acoustic features measures in the proposed analysis.

Table 4. Results observation of acoustic entropy measures in the proposed analysis

S.No	**Entropy Measures**	**Observation from the Obtained Results**
1	Approximate Entropy measure	However, self-similar patterns in the input signal are taken into account when assessing irregularity in the speech signal.
2	SVD Entropy measure	Decomposition of high-dimensional data using singular values to obtain the random measure
3	Spectral Entropy measure	Compiling the power spectral density function's random values
4	Ensemble entropy measures	The entropy of a network may also be defined. It's the logarithmic sum of all the networks' attractors.
5	Acoustic features measures	Emotions are detected by analysing the spoken signal's acoustic characteristics.

Figure 2 shows the graph of accuracy for overall emotion feature of acoustic phenomena. The feature sets, on the other hand, show variations.

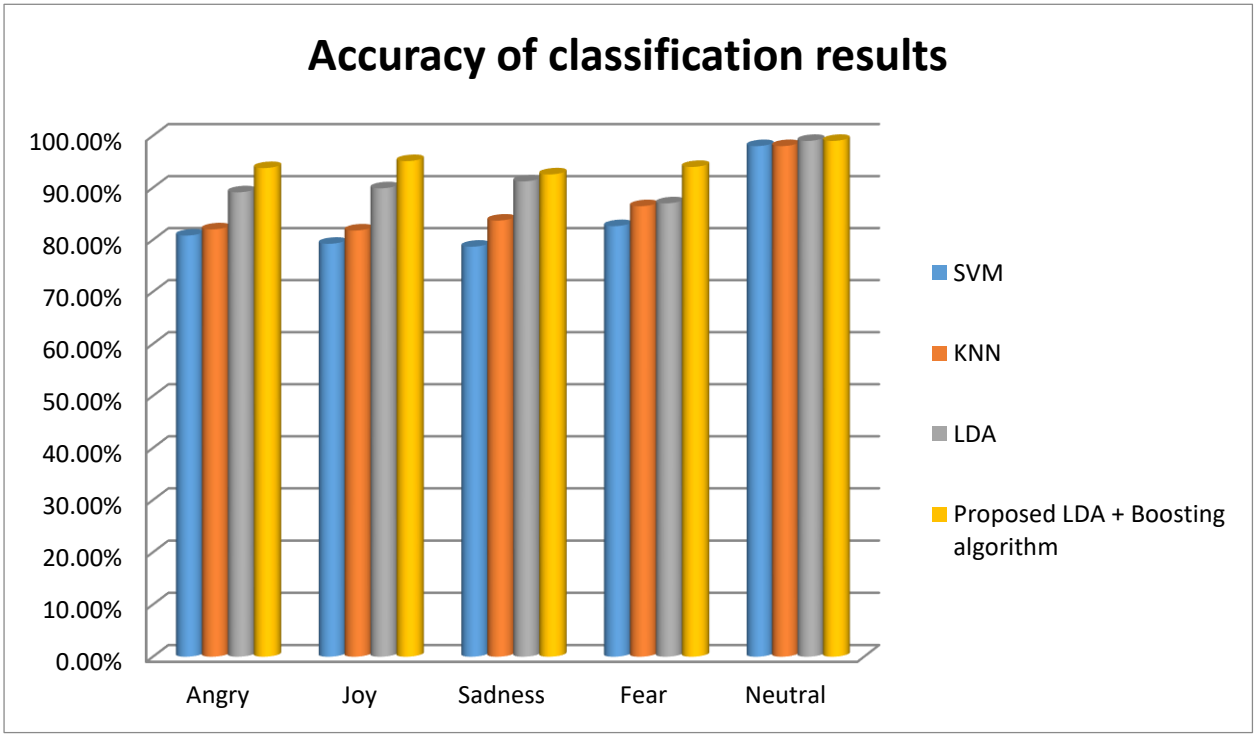

Figure 2. Graph of accuracy for emotion feature based acoustic phenomena

The proposed algorithm reaches highest possible accuracy for neutral emotion features that is showed in figure 2. This work considers the fact that no music genre has a clearly defined border and that the bounds of perception might be arbitrary at the best.

6. Conclusion

The validation of a collection of acoustic characteristics based on emotions that have been successfully recognized is presented in this study using a variety of external circumstances. Using the acoustic qualities provided, researchers were able to identify all eight emotions studied, including the fear emotion, which was previously thought to be a challenge. Prosodic and spectral characteristics are used in the suggested technique to boost the total discriminating power of the features, thereby enhancing the classification accuracy. When the same hybrid machine learning algorithms were applied to multiple feature sets, they were able to reveal how each collection of features performed when it came to their ability to discriminate. A state-synchronous model that doesn't suit the phoneme-level approach to fractal

features will be of interest in the future for statistical modelling for the fusion of numerous feature cues.

References

[1] Thakur, Amrita, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha, and Subarna Shakya. "Real Time Sign Language Recognition and Speech Generation." Journal of Innovative Image Processing 2, no. 2 (2020): 65-76.

[2] Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W,Sturge-AppleM(2017) Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. Int J Speech Technol 20(1):27–41. https://doi.org/10.1007/s10772-016-9364-2

[3] Tesfamikael, Hadish Habte, Adam Fray, Israel Mengsteab, Adonay Semere, and Zebib Amanuel. "Construction of Mathematical Model of DC Servo Motor Mechanism with PID controller for Electric Wheel Chair Arrangement." Journal of Electronics 3, no. 01 (2021): 49-60.

[4] Patel P, Chaudhari AA, Pund MA, Deshmukh DH (2017) Speech emotion recognition system using gaussian mixture model and improvement proposed via boosted gmm. IRA Int J Technol Eng (ISSN 2455-4480) 7(2 (S)):56–64

[5] Sathesh, A. "Metaheuristics Optimizations for Speed Regulation in Self Driving Vehicles." Journal of Information Technology and Digital World 2, no. 1 (2020): 43-52.

[6] Lanjewar RB, Mathurkar S, Patel N (2015) Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques. Procedia Comput Sci 49:50–57. https://doi.org/10.1016/j.procs.2015.04.226

[7] Manoharan, Samuel. "An improved safety algorithm for artificial intelligence enabled processors in self-driving cars." Journal of Artificial Intelligence 1, no. 02 (2019): 95-104.

[8] Partila P, Tovarek J, Voznak M (2016) Self-organizing map classifier for stressed speech recognition, p 98500A. https://doi.org/10.1117/12.2224253

[9] Anand, C. "Comparison of Stock Price Prediction Models using Pre-trained Neural Networks." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 3, no. 02 (2021): 122-134.

[10] Yang N, Dey N, Sherratt RS, Shi F (2020) Recognize basic emotional statesin speech by machine learning techniques using mel-frequency cepstral coefficient features. J Intell Fuzzy Syst.https://doi.org/10.3233/jifs-179963.

[11] Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3, no. 2 (2021): 95-113.

[12] Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117. https://doi.org/10.1007/s10772-011-9125-1

[13] Sungheetha, Akey, and Rajesh Sharma. "Transcapsule model for sentiment classification." Journal of Artificial Intelligence 2, no. 03 (2020): 163-169.

[14] Sailunaz K, Dhaliwal M, Rokne J, Alhajj R (2018) Emotion detection from text and speech: a survey. Soc Netw Anal Min 8(1):28. https://doi.org/10.1007/s13278-018-0505-2

[15] Tripathi, Milan. "Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM." Journal of Artificial Intelligence 3, no. 03 (2021): 151-168.

[16] Wieman, M.; Sun, A. Analyzing Vocal Patterns to Determine Emotion. Available online: http://www.datascienceassn.org/content/analyzing-vocal-patterns-determine-emotion

[17] Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.

[18] Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. Affect. Comput. **2015**, 7, 190–202.

[19] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier,W.F.;Weiss, B. A Database of German Emotional Speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005.

[20] Busso, C.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. Lang. Resour. Eval. **2008**, 42, 335.

[21] Livingstone, S.R.; Russo, F.A. 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 **2018**, 13, e0196391.

[22] Kerkeni L, Serrestou Y, Raoof K, MbarkiM,Mahjoub MA, Cleder C (2019) Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. Speech Commun 114:22–35.

[23] Chourasia, Mayank, Shriya Haral, Srushti Bhatkar, and Smita Kulkarni. "Emotion recognition from speech signal using deep learning." Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020 (2021): 471-481.

[24] Al-azani, Samah Ali, and C. Namrata Mahender. "Rule Based Part of Speech Tagger for Arabic Question Answering System." In International Conference on Communication, Computing and Electronics Systems, p. 733.

[25] Krishna, Akhila, Satya prakash Sahu, Rekh Ram Janghel, and Bikesh Kumar Singh. "Speech Parameter and Deep Learning Based Approach for the Detection of Parkinson's Disease." In Computer Networks, Big Data and IoT, pp. 507-517. Springer, Singapore, 2021.

[26] Gulati, Savy. "Comprehensive review of various speech enhancement techniques." In International Conference On Computational Vision and Bio Inspired Computing, pp. 536-540. Springer, Cham, 2019.

[27] Kalamani, M., M. Krishnamoorthi, R. Harikumar, and R. S. Valarmathi. "Swarm Intelligence Based Feature Clustering for Continuous Speech Recognition Under Noisy Environments." In International Conference On Computational Vision and Bio Inspired Computing, pp. 1248-1255. Springer, Cham, 2019.

[28] Akçay MB,O˘guzK(2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76. https://doi.org/10.1016/j.specom.2019.12.001

[29] Vijayakumar, T., Mr R. Vinothkanna, and M. Duraipandian. "Fusion based Feature Extraction Analysis of ECG Signal Interpretation–A Systematic Approach." Journal of Artificial Intelligence 3, no. 01 (2021): 1-16.

[30] Anttonen J, Surakka V (2005) Emotions and heart rate while sitting on a chair. In: Proceedings of the SIGCHI conference on Human factors in computing systems—CHI '05, ACM Press, New York, New York, USA, p 491.

[31] Sathesh, A. "Computer Vision on IOT Based Patient Preference Management System." Journal of Trends in Computer Science and Smart Technology 2, no. 2 (2020): 68-77.

[32] Davis, S.B., Mermelstein, P., 1980. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" IEEE Trans. Acoust., Speech, Signal Process. 28 (4), 357–366.

Author's biography

Subarna Shakya is currently a Professor of Computer Engineering, Department of Electronics and Computer Engineering, Central Campus, Institute of Engineering, Pulchowk, Tribhuvan University, Coordinator (IOE) , LEADER Project (Links in Europe and Asia for engineering, education, Enterprise and Research exchanges), ERASMUS MUNDUS. He received MSc and PhD degrees in Computer Engineering from the Lviv Polytechnic National University, Ukraine, 1996 and 2000 respectively. His research area includes E-Government system, Computer Systems & Simulation, Distributed & Cloud computing, Software Engineering & Information System, Computer Architecture, Information Security for E-Government, and Multimedia systems