FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

SPEECH EMOTION RECOGNITION: APPLICATION IN DISTANCE LEARNING EDUCATION

Dahiru TANKO

Master's Thesis

DEPARTMENT OF DIGITAL FORENSIC ENGINEERING

FEBRUARY 2022

FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

Department of Digital Forensic Engineering

Master's Thesis

SPEECH EMOTION RECOGNITION: APPLICATION IN DISTANCE LEARNING EDUCATION

Author
Dahiru TANKO

Supervisor
Doç Dr. Sengul DOGAN

FEBRUARY 2022
ELAZIG

FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

Department of Digital Forensic Engineering

Master's Thesis

Title:	Speech Emotion Recognition: Application in Distance Learning Education
Author:	Dahiru TANKO
Submission Date:	03 January 2022
Defense Date:	07 February 2022

THESIS APPROVAL

This thesis, which was prepared according to the thesis writing rules of the Graduate School of Natural and Applied Sciences, Fırat University, was evaluated by the committee members who have signed the following signatures and was unanimously approved after the defense exam made open to the academic audience.

		Signature
Supervisor:	Doç Dr. Sengul DOGAN Firat University, Faculty of Technology	Approved
Chair:	Doç. Dr. Türker TUNCER Fırat Universitesi, Faculty of Technology	Approved
Member:	Dr. Öğr. Üyesi Mehmet BAYGIN Ardahan University, Faculty of Engineering	Approved

This thesis was approved by the Administrative Board of the Graduate School on

..... / ..... / 20 .....

Signature

Prof. Dr. Kürşat Esat ALYAMAÇ
Director of the Graduate School

DECLARATION

I hereby declare that I wrote this Master's Thesis titled " Speech Emotion Recognition: Application in Distance Learning Education" in consistent with the thesis writing guide of the Graduate School of Natural and Applied Sciences, Firat University. I also declare that all information in it is correct, that I acted according to scientific ethics in producing and presenting the findings, cited all the references I used, express all institutions or organizations or persons who supported the thesis financially. I have never used the data and information I provide here in order to get a degree in any way.

07 February 2022

Dahiru TANKO

PREFACE

There has been a surge in the adoption of online learning systems as a result of the covid outbreak. This research is geared toward the application of Speech emotion Detection in detecting an emotion of a lecturer during a teaching process in an online learning environment, in other to measure the performance of the educators. This will ensure that the quality of learning is maintained in a virtual classroom. It is a big step towards improving educational technology.

This research is an aspect of speech processing that is limited to emotion detection from speech in a virtual classroom environment, as such, it requires a dataset of audio lessons from this platform for optimum results. This however is difficult due to some limitations in the education technology that will help in the capturing of the speeches. We collected the dataset from participants that were willing to give these lessons from a prepared lesson material, expressing the required emotions in their utterances. The result is promising and can effectively be integrated into an online learning system.

I would like to express my profound gratitude to my supervisor, **Associate Professor Sengul DOGAN** for her immeasurable support and guidance throughout my study and thesis. I will also like to extend my gratitude to **Associate Professor Turker TUNCER** for all his guidance and support. My profound gratitude to my family, most especially my mother **Mrs. Habibah TANKO** for all her prayers and support.

I would like to extend my profound gratitude to the Nigerian **National Information Technology Development Agency (NITDA)** for granting me a scholarship to embark on this masters' program.

Dahiru TANKO
ELAZIG, 2022

CONTENTS

ABSTRACT

Speech Emotion Recognition: Application in Distance Learning Education

Dahiru TANKO

Master's Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences

Department of Digital Forensic Engineering

February 2022, Page: x + 53

Affective computing is a branch of artificial intelligence that tries to pass the innate human capabilities of emotional intelligence to machines to enhance a smooth interaction between humans and computer systems. Speech emotion recognition is an essential aspect of affective computing and plays a significant role in designing systems and machines that recognize, analyze, interpret, and simulate human emotional states. In this project, the concept of speech emotion recognition is being integrated into a distance learning system to investigate the lecture delivering performance via a distance learning platform. To archive this, the performance is classified into three categories: interesting, Neutral, and boring, depending on the lecturer's emotional state. The project is implemented using intelligent machine learning techniques to recognize and interpret a lecturer's emotional state. It is carried out in four major stages of data preprocessing, feature extraction, feature selection, and classification. Our model adopts a comprehensive feature generation approach that utilizes a shoelace graph pattern as a local feature generator alongside tunable Q wavelet transform (TQWT). The best four feature vectors from the feature generation stage are selected and merged to obtain the final feature vector. After that, we applied an NCA method at the feature selection stage to select 512 most discriminative features. We then perform the classification using the SVM classifier. Our proposed network performed well, giving us an accuracy of 96.41% when applied on the Turkish dataset and 94.97% when applied on the English dataset.

Keywords: Speech emotion recognition, Distance learning, Shoelace pattern, Machine learning; Artificial intelligence.

ÖZET

Uzaktan Eğitim Uygulamalarında Konuşma Duygu Tanıma

Dahiru TANKO

Yüksek Lisans Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Adli Bilişim Mühendisliği Anabilim Dalı

Şubat 2022, Sayfa: x + 53

Duyguları tanımlama/hesaplama, insanlar ve bilgisayar sistemleri arasında sorunsuz bir etkileşimi geliştirmek için duygusal zekanın doğuştan gelen insan yeteneklerini makinelere aktarmaya çalışan bir yapay zeka dalıdır. Konuşmadan duygu tanıma, duygusal hesaplamanın önemli bir yönüdür ve insan duygusal durumlarını tanıyan, analiz eden, yorumlayan ve simüle eden sistem ve makinelerin tasarlanmasında önemli bir rol oynar. Bu projede, konuşmada duygu tanıma kavramı, bir uzaktan öğrenme platformu aracılığıyla performans sunan dersi araştırmak için bir uzaktan öğrenme sistemine entegre edilmektedir. Bunu arşivlemek için performans üç kategoride sınıflandırılır: eğitmenin duygusal durumuna bağlı olarak ilginç, Nötr ve sıkıcı. Proje, bir öğretim görevlisinin duygusal durumunu tanımak ve yorumlamak için akıllı makine öğrenimi teknikleri kullanılarak uygulanmaktadır. Veri ön işleme, özellik çıkarma, özellik seçimi ve sınıflandırma olmak üzere dört ana aşamada gerçekleştirilir. Modelimiz, ayarlanabilir Q dalgacık dönüşümünün (TQWT) yanı sıra yerel öznitelik üreteci olarak bir ayakkabı bağı grafik desenini kullanan kapsamlı bir öznitelik oluşturma yaklaşımını benimser. Nihai öznitelik vektörünü elde etmek için öznitelik oluşturma aşamasından en iyi dört öznitelik vektörü seçilir ve birleştirilir. Daha sonra, en ayırt edici 512 özniteliği seçmek için öznitelik seçim aşamasında komşuluk bileşen analizi (NCA) yöntemi kullanılmıştır. Ardından destek vektör makinesi (SVM) sınıflandırıcısını kullanarak sınıflandırmayı gerçekleştiriyoruz. Önerilen yöntem iyi performans gösterdi ve bize Türkçe veri setine uygulandığında %96,41 ve İngilizce veri setine uygulandığında %94,97 doğruluk sağladı.

Anahtar Kelimeler: Konuşmadan duygu tanıma, Uzaktan Eğitim, Ayakkabı bağı örüntüsü, Makine Öğrenmesi, Yapay Zeka

LIST OF FIGURES

Page

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS

Symbols

α : Signal Scaling Parameter (Low-pass Sub-Band)
β : Signal Scaling Parameter (High-pass Sub-Band)

Abbreviations

ShoePatNet23 : Shoelace Pattern Network 23
SER : Speech Emotion Recognition
NCA : Neighborhood component Analysis
TQWT : Tunable Q-Factor Wavelet Transform
DWT : Discrete Wavelet Transforms
HCI : Human-Computer Interaction
Ed-tech : Education Technology
Acc : Accuracy
SVM : Support Vector Machine
CNN : Convolutional Neural Network
MLP : Multilayer Perceptron
BLSTM : Bi-directional Long Term Short Term memory
HMM : Hidden Markov Model
MLR : Multivariate Linear Regression
GMM : Gaussian Mixture Model
FMLLR : Feature-space Maximum Liklihood Linear Regression
MEDC : Mel-energy Spectrum Dynamic Coefficient
PNN : Probablistic Neural Network
VQ : Vector Quantization
ZCR : Zero Crossing Rate

1. INTRODUCTION

Speech emotion recognition (SER) is the study and analysis of speech signals via a collection of carefully selected methodologies to detect, process, and classify the embedded emotions in a speech signal. It is a research area that has seen a rapid increase in research interests due to its numerous area of applications, most especially, its application in enhancing human-computer interactions (HCI). Emotion is a conscious reaction subjectively experienced as a strong feeling, usually directed towards a specific object and typically accompanied by physiological and behavioral changes in the body [1]. In simple terms, it is a state of feeling. Humans display one form of emotion or the other at any point in time depending on their inner state or mood. Their emotional state often determines how they act and react to the people and things around them. It can be expressed in one's physical appearance such as in the facial and general body reaction and can also be embedded in the words we speak. This makes the voice an important means of communication as it has been shown to not only convey meanings but also emotional state. It is possible to identify these emotions through spoken words because when a word is spoken, the tone and the pitch of the sound alongside other speech features convey the embedded emotion in it. It is evident that when people are happy, they tend to speak differently as to when they are sad, or bored. Happy is a positive emotion that is characterized by excitation features of speech [2]. Happy voices are generally known to be loud with considerable variability in loudness, have a high and variable pitch, and also a higher speech rate [1]. Putting these three features of speech together, we created a new speech emotion dataset with three different classes namely; positive, neutral, and negative which is used alongside our novel speech emotion recognition model to classify the lectures which are delivered via a distance learning platform. Given the emotional state of the lecturer, our model will be able to tell if the lecture is interesting, neutral, or boring. This is because there is a direct correlation between one's emotional state and his reaction to his environment and that emotion plays a crucial role in conveying ideas during communication [3]. A lecturer in a happy mood gives a speech with high excitation features and would spark the learning atmosphere and enhance the students learning abilities. But when the lecturer is bored, his speech has low excitation features and could impact the learning abilities of the students in a negative manner. Many studies have been carried out on speech emotion recognition, that detects and classify the basic and complex emotions using some popular speech emotion datasets and or electroencephalogram (EEG) datasets [4-8], But this research focuses on the application of speech emotion recognition in distance learning, we will be focusing on those emotions that are related to academics and that influences learning in an academic environment. These emotions are called academic emotions.

Research has shown that students' learning and achievement are greatly influenced by these academic emotions [9]. This influence has been validated concerning how academic emotions

motivate the students towards learning. For instance, learners are more likely to spend their time and resources if the learning environment is laden with activities that are enjoyable and interesting rather than anxiety-laden or boredom-inducing [10]. Students' interest in learning materials is triggered and maintained by emotions [11]. Emotions also stimulate the more cognitive aspects of learning, as they can induce diverse states of information processing and problem-solving in a learner [12], and influence or hinder students' self-regulation of learning [13]. The impact which academic emotions have on students' performance and achievement is associated with cognitive and motivational mechanisms such as reasoning resources, learning motivation, and methods of learning [14, 15]. Positive emotions like enjoyment, hope, and pride certainly possess a positive influence on motivation, the use of easy learning approaches and self-regulation, and the availability of cognitive resources for task engagement [13]. As opposed to this, negative emotions impede motivation and trigger the requirement for the adoption of more inflexible strategies, such as simple rehearsal and algorithmic procedures. In particular, negative effects such as anger, anxiety, and boredom are thought to impair cognitive resources and self-regulation [13, 16].

Most studies have implemented SER using deep learning techniques achieving a good level of accuracy. Since deep learning techniques are known to have computational and time complexities, we developed a handcrafted method of feature generation technique alongside a classical machine learning technique for the classification of emotions. Our SER network adopts a comprehensive feature generation approach that modeled a shoelace framework hence the name shoelace pattern. It is a local feature generation method that generates low-level features. To further enhance the feature generation ability of our network, a tunable Q wavelet transform (TQWT) is applied to decompose the speech signal into sub-bands where the shoelace pattern is used to extract features from both the raw speech and the sub-band signals. The best four feature vectors are selected and merged to obtain a final feature vector. Neighborhood component analysis (NCA) method has been used to select the 512 most informative features which are used in the classification stage with an SVM classifier to classify the emotions.

1.1. Problem Statement

There has been an increase in the adoption of the distance learning education system among many learning institutions. However, one major problem faced in this system is how to effectively measure the effectiveness of the online learning system in other to ensure the quality of learning.

In line with the inquiries into the role of teachers' emotion, enthusiasm, principles of teaching, and classroom management in learning , we try to develop an intelligent machine learning technique that will classify and rate a lesson given in a virtual classroom based on the emotions displayed by the teachers during the teaching process.

1.2. Purpose

The outbreak of the covid-19 pandemic has opened the door for the wider adoption of distance learning systems among many learning institutions [1]. As a result, learners are made to take lectures and learn from the comfort of their homes. This ensures the continuation of learning in the wake of the lockdown that governments introduced to limit physical contacts among people to curtail the spread of the covid-19 virus. This has changed the education system dramatically, with the distinctive rise of e-learning, whereby teaching is undertaken remotely and on digital platforms. As a result of this, many learning institutions have come to realize other benefits of adopting the distance education learning system, such as cost reduction (for the students and management) and convenience in terms of the flexibility it offers. Furthermore, research suggests that online learning has been shown to increase retention of information, and take less time, meaning the changes coronavirus have caused might be here to stay. Therefore, these institutions are likely to continue embracing this learning system either in totality or partially in the post-coronavirus era. Even before the emergence of the coronavirus, there was already a surge in growth and adoption in education technology (ed-tech), with global investments in education technology exceeding eighteen billion U.S dollars as of 2019. The overall market for online learning has been projected to reach about $350 Billion by 2025. Whether it is language apps, virtual tutoring, video conferencing tools, or online learning software, there has been a significant surge in usage since COVID-19.

Hence, there is a need to put in place a measure to ensure the quality of learning via the e-learning system. One such possible measure is by checking and analyzing the performance of lecturers or teachers during the lecture. This will help to improve upon to meet the necessary standards, format, and quality of online courses as well as virtual classrooms. One way to archive that is to measure the performance of the one presenting the lecture (instructor, teacher, and or lecturer) In this work we choose to do that by analyzing the emotional state of the lecturer while giving a lecture or lesson to determine if he displays the appropriate affects that fosters learning since our emotion at a particular point in time determines or contributes to our mood and can affect how we interact with others then. given the emotional state, it is possible to tell if the lecture is interesting or not and if it is neutral. Hence, we will be able to measure the quality of the learning taking place via these platforms.

1.3. Hypothesis

This research is carried out to verify the hypothesis that, When viewed as a pattern recognition problem, speech emotion recognition can be implemented using machine learning

techniques to detect and classify embedded emotions in a speech signal, and that it can be effectively employed in a learning environment to monitor the quality of learning.

1.4. Thesis Structure

The thesis is written and presented in the format below.

Chapter one is an introduction to the study. It is a brief statement about the importance of the research subject and a brief explanation of the method and the result we obtained.

Chapter two present the studies in the literature that have tried to tackle the same subject either directly or indirectly. In this chapter, we try to present the results of these other studies as well as their methods and materials.

Chapter three discuss in detail, the theories behind some popular concepts that were adopted in this study from speech generation, features of speech, emotions, and how it is related to learning and cognition. We also talked about voice emotional cues that were employed in generating the dataset for this research.

In chapter four, we discuss in detail the materials we used for this experiment most especially the dataset. This is the heart of this project as it explains in detail every aspect of the developed model. It also explains the algorithms used and some mathematical backing of some theories and concepts that were employed in the model.

In chapter five, we present the result and findings of our model. We discuss the model performance on the different datasets that were used using tables and charts.

Chapter six is a conclusion from our experiment and highlights the contribution of our research to the body of knowledge. And we further put forward a recommendation as well.

2. RELATED STUDIES

The concept of speech emotion recognition has continued to receive attention because of humans' quest to develop an emotionally aware machine to make human-machine interaction more human-like. It has always been the central goal of artificial intelligence to create a system that models a human in character and appearance. To achieve this level of intelligence, a precise and complete understanding of human language (both in context and paralinguistic properties) will always be the core goal of AI. Emotion and cognition are major components of this. therefore research on speech emotion recognition has gained attention continuously over the last decades. Some models on speech emotion recognition have been suggested in the past. Tuncer et al [4] postulated a model for accurate speech emotion recognition tasks. Their model uses a feature generation technique called the Twine shuffle pattern (Twine-Shuff-pat) and a TQWT technique to extract high-level features for the classification of emotion. Dias Issa et al [5] in their model used a deep learning framework. Their model architecture extracts mel frequency cepstral coefficient (MFCC), Chromagrams, Mel-scale spectrograms feature and uses one-dimensional convolutional neural network (CNN) for classification. they were able to achieve 95.51% for the berlin database of emotional speeches (EMO-DB) dataset. Shuzhen et al [17] proposed a spatiotemporal and frequential cascaded attention network to tackle the challenges of extracting emotional features from a long utterance. J. Ancilin et al [18] in their work choose to extract a Mel frequency cepstral coefficients from the magnitude spectrum of the signal instead of the energy spectrum with the exclusion of discrete cosine transform to extract a Mel Frequency Magnitude Coefficient. After classifying with SVM, they achieved an accuracy of 95.25% for the Urdu database and 81.50%, 64.31%, 75.63%, for Berlin Database of Emotional Speech (EMO-DB), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE) datasets respectively. Qiupu et al [19] proposed a dual-level architecture which they called dual-attention bi-directional short term, long term memories (BLSTM) for SER. They also introduced a new data preprocessing technique of linear interpolation and decimation and tested their model on the interactive emotional dyadic motion capture database (IEMOCAP) dataset and obtained an accuracy of 70.29%. Londhe et al [20] in their work of speech emotion recognition on the Indian dataset presented different machine learning techniques to Speech concepts. They employed the MFCC features from the speech signal and carried out classification tasks with SVM, Hidden Markov Model (HMM) and artificial neural network (ANN). The SVM classifier performs better with an accuracy of 90.60% after evaluation with various paradigms in the machine learning classifier. In a study by Kerkeni et al [21], they presented a Methods and Case Analysis of feature extraction methods for Speech Emotion Recognition tasks with MFCC feature. They performed the classification task using SVM and Multivariate Linear Regression (MLR). From the result of their

experiment, the best result was obtained when they combined the mel frequency cepstral coefficient (MFCC) feature with Mass Spectrometry features, the accuracy was 90%. A neural network of Multilayer Perceptron is employed in another SER task by Cai et al [22]. They employed the Gaussian mixture model (GMM), feature-space maximum likelihood linear regression (FMLLR), and MFCC features. they used multilayer perceptron (MLP) to carry out the classification task. In Kerkeni et al [23], and Automatic SER Using Machine Learning was proposed that uses MFCC and Linear Predictive Coding (LPC) as the features for the task. The authors used MLR, recurrent neural network (RNN), and SVM classifiers of which SVM gave an experimental accuracy of 86%. The SVM has the advantage that it required a lesser amount of data as compared to RNN. In Pan et al [24], they presented speech emotion recognition research that utilized SVM as a classifier and LPCC, MFCC, and Mel Energy Spectron Dynamic Coefficient (MEDC) features. The highest experimental result from this study is 90% using the combination of MFCC + MEDC + Energy + Pitch features. A probabilistic neural network (PNN) was used by Palo et al [25] for SER. Here the strategies used are the sub-band spectral to Vector Quantization (VQ) for feature extraction where VQ is computationally less complex. 90% accuracy was attained with the PNN and it proved to be the fastest classifier as well (for this task) and is less complex. Javidi et al [26] presented an SER study using C5.0, NN, and SVM Classification algorithm. The features used here are Zero cross-rate (ZCR) and MFCC. The system reached 89% classification precision. The NN improves the systems' response-ability to uncertain circumstances. Palo et al [27] put forward an Emotion recognition study using Multilayer Perceptron and GMM for the Oriya Language. They employed many speech features alongside MFCC, Perceptual Linear Prediction (PLP), and LPC. After classification, MLP offers the highest accuracy with 87% accuracy. The combination of PLP and LPC always yield better result and at the same time fast with a lower frequency of operation. In another study, Motamed et al [28] came up with an emotion detection model that is based on a revised brain affective learning standard. It integrates the concepts of the brain computing interface (BCI). MLP and Adaptive Neuro-Fuzzy Inference Method (ANFIS) were used as the classifiers with MFCC being used as the feature. MLP attained 84% precision accuracy of a human speech emotion. ANFIS has found extensive use in many platforms due to its excellent detection, simulation, and conflict resolution capabilities and has been implemented in many multiple platforms. In Chiou et al [29], an SVM was adopted and fed with the ZCR and MFCC features of the speech for the SER task. They achieved 85% accuracy. They also employed Principal Component Analysis (PCA) for dimensionality reduction. The ZCR feature helps improve the efficiency of the model. Jacob et al [30] suggested the use of logistic regression (LR) and decision trees (DT) as classifiers to model an emotion detection system. MFCC features were extracted and used for the classification. The decision tree gave classification accuracy of 93.63% while the highest accuracy that was obtained with logistic regression was 73%. Decision trees have less

computational complexities with fewer requirements elicitation or data preparation during pre-processing in comparison with other classification algorithms. A Comparative Neural Network Study for SER was conducted by Palo et al [31] using MFCC, PLP, and Linear Prediction Cepstron Coeficient (LPCC) Vector Quantization (VQ) features. an accuracy of 83% was attained with an NN whereas MLP achieve an accuracy of 80%. The neural network handled the noise and missing data present in the data better. The author proposed in another study [32] the use of LPC and VQ features with a Radial Base Function Neural Network (RBFNN) as the classifier. The study introduced a powerful combination of design strategies combining the feature sets of (LP VQC and pH VQC) which increases the mean precision and accuracy to 90.55%. Shaqra et al [33] developed a model for the recognition of emotion from Age- and Gender-based Voice built-in classifiers using Hierarchical ML Models. EGeMAPS and MFCC have used features. The Hierarchical MLP Classifier improves the efficiency of the model and increases the accuracy of the model to 74% after the models are integrated. Prasomphan et al [34] in their research of SER used LPCC and MFCC spectral features. they carried out a classification comparison with ANN, SVM, and GMM, in which SVM gave a better prediction accuracy in different circumstances. It gave an accuracy of 85%. Chen et al [35] postulate a classification model for emotion detection systems. In their system, they employed features and applied an SVM, ANN for classification tasks. Before the classification, a PCA was used to reduce the dimension of the extracted MFCC features which improved the model outcome and accuracy. Four different experiments were conducted using a combination of characteristics and classifiers. But SVM, Fisher combination came out on top with 85.6% accuracy. A Speaker-dependent speech emotion recognition was presented in a study by Dahake et al [36] using the MFCC feature with an SVM classifier. KNN and HMM classification algorithms were also used to compare the result with that of SVM. SVM provides 84% accuracy using autocorrelation and Average Magnetude Difference Function (AMDF) algorithms. To reduce the computational cost and the complexity of the model, the AMDF values of the speech signal structure are converted into one-bit signals. In another research by Rajisha et al [37] on Malayalam Language Emotion Detection Framework for performance analysis using ANN and SVM classifiers. They also used the MFCC features and the pitch. The result analysis indicates that emotion detection with ANN yields a higher accuracy (88.4%) as compared to SVM. ANN models how the human brain functions as the basis for designing algorithms to model dynamic patterns and predictive problems dividing the audio into a segment of small frames thereafter detecting the pitch at each frame. Pan et al [24] proposed a study of SER using an SVM classifier in combination with various features like F0, ZCR, LPC, and MFCC. The MFCC has the best consistency in giving the best accuracy with linear kernel on all databases. It achieved 89.80%, 93.57%, and 98% accuracy on Berlin, Japan, and Thai emotion repositories respectively. It also gave an accuracy of 71.8 And 88.7% for the qualified and test data collection in RBF(Kernel). The SVM is useful in

conditions where the number of measurements is more than the number of samples and is relatively stable in the memory. Yu et al [38] presented an SER model using an Optimized SVM classifier. MFCC feature is employed for this task. RBF kernel functions were used to get the Optimized SVM, which provides 88.75% accuracy. SVM classifier is very useful in a high dimensional space. Chenchah et al [39] presented a speech detection task in a noisy atmosphere. Spectral conditions Subtraction, wavelet transforms and Minimum Mean Squared Error (MMSE) and MFCC features were used with HMM classifier. The technique discards the comparison of error rate with SER additive noise.

In a study by Likitha et al [40] on human Speech emotion detection with an MFCC feature of the speech. The classification algorithm of SVM and HMM was used. This model gave a recognition rate of 80%.

Kamaruddin et al [41] proposed a model for analysis of driver behavior through speech emotion detection. The feature used here is MFCC and the classifier is an MLP classifier. Adaptive network-based fuzzy interface system, Generic Self organizing of fuzzy neural networks were integrated into the classifier during implementation. 70% accuracy was obtained. Bandela et al [42] presented a study on Stressed speech emotion recognition. A GMM learning algorithm was used as a classifier with T-MFCC, 93.3% prediction accuracy was attained which is comparatively better than the MFCC feature. Iliou et al [43] introduced Comparative Analysis of SVM-MLP-PNN classifiers for Speech Emotion Recognition. Here the MFCC, Pitch are used as features to the classifiers. This analysis shows that PNN has a better accuracy of approximately 94%. Peerzade et al [44] proposed a study of SER. Tactic and SVM is one of the better classifiers than MLP. SER application is in demand in the area of Psychological diagnosis, robotics, mobile speech recognition technology, emotion identification in call centers, voice mail delivery, which are few mentioned here. In Khanchandani et al [45] they proposed an emotions identification model using MLP and a generic feed-forward neural network. The purpose of the MLPNN is to extract prosodic features while GFFNN is used for classification. The following overall accuracy was achieved with MLPNN, 93%; and GFFNN, 90%. In Reddy et al [46] a model was presented that is a fusion of many algorithms. They utilized MFCC, DWT features, and KNN, SVM classifiers. The performance of the model was greatly improved when approximating the frequency of time information to 94%. The Fusion-based algorithm gave a good performance for this task.

Studies have also been carried out on the application of affective computing in the distance education system. Most of the studies however focus on the emotional state of the learner. Maryam Imani et al [47] presented a survey of SER in an e-learning system where they compared different emotion recognition methods. Ken Chen et al [48] present a study on speech emotion recognition in e-learning based on eight different emotions using a neural network. They achieve 50% classification accuracy. Arindam et al [49] present a study on the implementation of an effective e-

learning strategy based on biophysical signals of facial expression for the detection of learner's emotions. Cen, Ling et al [50] postulate an SER model that explored emotion recognition from continuous speech to come up with a real-time speech emotion recognition system. Their SER system comprises voice activity detection, implemented fully using machine learning techniques that incorporate speech segmentation, speech signal pre-processing, feature generation, emotion classification, and statistical analysis of affective frequency. They tested their model with both pre-recorded datasets and real-time recordings conveyed in four distinct affective groups. They achieved average accuracies of 90% and 78.78% in the two experiments, respectively. headings

3. THEORETICAL BACKGROUND

This section describes the theories behind the different concepts that have been used in this research.

3.1. Speech Signal

Speech is a human vocal communication that uses phonetic combinations of vowel and consonant to form a sound of words[1]. It is created at the vocal cords, travels through the vocal tract, and is produced at the speaker's mouth. It gets to the listener's ear as a sound wave which is then decoded by the listener based on the semantic and syntactic characters of the words. The voice or vocal signal constitutes a particularly important means of communication. Vocal signals have been shown to convey not only relatively enduring features like age and gender but also a wide range of transitory states such as health and power [51]. It has been proposed that the human voice also conveys emotional states, each characterized by a unique acoustic profile[52-54]. During communication, speakers perform many different intentional and unintentional speech acts, e.g., persuading, asking, informing, declaring, directing, and can also use enunciation, intonation, degrees of loudness, tempo, and other non-representational or paralinguistic aspects of vocalization to convey meaning. In their speech speakers also unintentionally communicate many aspects of their social position such as sex, age, place of origin (through accent), physical states (alertness and sleepiness, vigor or weakness, health or illness), psychic states (emotions or moods), physio-psychic states (sobriety or drunkenness, normal consciousness, and trance states), education or experience, and the like. When a listener hears this sound, he/she synthesizes this sound to extract this information through active thinking as the speaker must draw upon past experiences with the speaker symbol and paralinguistic agents to align his thinking with that of the speaker[55] in other to gain a true perception of the message being passed across. It is very important to learn from our perception of speech sound in other to be able to develop systems or machines that truly understand human language. Speech is perceived as an audio signal which is a form of sound with a frequency range between 20 to 20,000 Hz that corresponds to the lower and upper human hearing limit. It is an analog signal and also requires processing for a speech-related task as most of these tasks require signals in a digital state. Speech signal processing is, therefore, regarded as a special case of signal processing

3.1.1. Features of Speech

Every speech processing application employs certain properties or features of speech signals in other to achieve its target goal. Audio features are the characteristic of sound or an audio signal

that can be extracted and fed into statistical or machine learning algorithms to build intelligent systems. Audio classification, speech recognition, automatic music tagging, audio segmentation, and source separation, audio fingerprinting, audio denoising, music information retrieval, are some of the audio applications that use such features include and more. Audio features are predominantly categorized based on: level of abstraction, temporal scope, musical aspect, signal domain, and machine learning approach. These all encapsulate the statistics of sound.

Level of abstraction: This is an audio feature categorization method that applies to musical signals and not on every audio. The audio is categorized into (i) high-level, (ii) mid-level, and (iii) low-level signals.

Temporal aspect: This audio feature categorization method applies to all audio signals and not just musical audio. It categorizes features into (i) instantaneous, (ii) segment level, and (iii) global.

Musical aspect: This categorizes audio features based on acoustic properties that include beat, rhythm, timbre (color of sound), pitch, harmony, melody, etc.

Signal domain: It is the most useful signal categorization method for performing machine intelligence tasks. It categorizes features into (i) time-domain features,(ii) frequency-domain features, and (iii) time-frequency domain features. The signal domain features consist of the finest or rather descriptive features for all kinds of audio signals.

As stated above, signal domain features are the most essential features for audio processing for machine learning tasks. There are two major categories of the signal domain features, these are temporal and spectral features. They both carry unique information about the speech signals and can be retrieved to aid the proper analysis of speech signals.

Temporal Features: These refer to those features of the speech signal that are retrieved in the time domain. They have a simple physical representation and are very easy to extract because they are usually extracted from the waveform of the raw audio. Some of the time domain features are zero-crossing rate, signal energy, amplitude envelope, minimum energy, etc.

Spectral features: These are features of the speech signal in the frequency domain. They are obtained after the transformation of the signal from the time domain to the frequency domain using a signal transformation method such as Fourier transform. Examples include band energy, fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. These features can be used to identify the notes, pitch, rhythm, etc.

Time-frequency features: These features combine both the time and frequency components of the audio signal. The time-frequency representation is obtained by applying the Short-Time Fourier Transform (STFT) on the time domain waveform. Spectrogram, Mel-spectrogram, and constant-Q transform are examples.

3.2. Emotion

There has not been an agreed-upon definition of emotion in the literature due to its complex and ambiguous nature[56, 57] despite several attempts by researchers to truly understand what it means. Though synonymous, it is often mistaken to mean the same thing with affect, feeling, and mood. And "Mood" is often subtler, long-lasting, less intensive, more in the background, giving the affective state of a person a tendency in a positive or negative direction [58]. The concept of emotion applies to all aspects of human existence[59], and its study has been on for centuries[12]. A comprehensive understanding of emotion is very vital in social psychology[60], and in artificial intelligence, as humans try to develop emotionally aware machines and systems. Furthermore, the recent emergence of cognitive neuroscience as an inspiration for understanding human cognition has highlighted its interaction with emotion[61, 62]. Though the definition of the concept of emotion might be quixotical, it will be very important to include the component of subjective experience, accompanying physiological responses, behavioral expression, and consequences[59]. Therefore, the term emotion can be referred to as a conscious reaction subjectively experienced as a strong feeling which is directed toward a specific object[63], typically accompanied by physiological and behavioral changes in the body. Simply put, it is a mental state of being. Several works in the literature have explored different ways in which emotionally valenced information can direct or influence attention[64-66], and influence decision processes[67]. Some researchers view emotion primarily as a psychobiological phenomenon[68] while others view them as primarily psychological[69], hence the concept varies from researcher to researcher.

3.2.1. Classes of Emotion

Despite the universality nature of affective states, there exists a constant argument over the concise nature and the number of basic emotions exhibited by humans. Some literature categorizes the basic emotions into five[70], others categorize them into six [71], and some categorize them into seven[72]. Darwin in his evolutionary work on emotion as a means of adaptation for survival was perhaps the first to systematically identify and categorize a comprehensive range of basic emotions in connection to the primitive instincts of survival[72]. His classification included over thirty different emotions that were further categorized into seven groups, merging similar emotions [73]. Since then, several emotion researchers have agreed on the existence of two types of human emotions: Primary and Secondary. Primary emotions are evoked by a stimulus, e.g., fear in sudden noise. Secondary or social emotions such as embarrassment, jealousy, guilt, pride are those that arise later in an individual's development when systematic connections are identified between primary emotions and categories of objects and situations [74]. Though the precise number of the basic emotion might be a subject of debate, the most and easily recognizable affective states, that

influences the categorization of other affective states are Happiness, Sadness, fear, anger, disgust, and surprise: as agreed upon by most researchers based on Darwin evolutionary work and are regarded as universal emotion since they cut across different cultures.

Some researchers came up with a different approach to classifying emotions, different from the conventional method of labeling. One of these approaches profers solutions towards the quantification of emotion in a dimensional space. An example is a circumplex model of emotional categorization in a spatial domain[75] deploying valence and arousal dimensions instead of emotion labels, projecting the user's affective state in a two-dimension emotional space. Another approach is one that referred to several indicators that identify the quality of an effect that are mostly used to measure emotions[76]:

- Arousal (activating/deactivating)
- Valence (positive/negative)
- Intensity (high–low)
- Duration (long-short)
- Frequency of its occurrence (frequent–occasionally)
- Time (retrospective like relief, actual like enjoyment, prospective like hope).

There is also an approach that perceives emotions as a state which are specific to a situation versus trait (apply to a broader context) that follow three different categories[77]:

- Core affect (moods like feeling blue),
- Emotional episodes (state emotions like sadness), and
- Affective tendencies (trait emotions like being depressed).

3.2.2. Emotion and learning

There is an increasing interest in and knowledge about the relationship between emotion and learning. There is rarely a learning process without emotions [76]. The concept of the significance of the close relationship of learning and emotion has already been pointed out by the early Greek philosophers like Aristotle, by renowned psychologists like Wilhelm Wundt and by pioneering educators like Maria Montessori, therefore, it is not at all new. Despite the glaring relationship between learning and emotion, not much is known about it. For a long time, learning was mainly analyzed in terms of cognitive or motivational aspects. Many learning theories have failed to acknowledge the affective processes for a long period. To gain a deeper insight into the complex area of learning they focused on cognition only. To answer the basic questions of how and why emotions influence the learning process, The interdependency of learning and emotion has to be analyzed theoretically and further investigated empirically. Several theories hold at least some empirical evidence [76]. The majority of these theories are deducted from an experimental study on mood induction, which studied the effects of mood on information processes. Mood induction

is a process of inducing a subject with a pleasant and unpleasant emotion for some time. This can be achieved by using personal memories, or by being confronted with positive or negative information, or by experiencing a positive or negative situation, and so on.

1. One of the most referenced theories of emotion and learning is called ‘mood-congruence-hypothesis’: It is based on the idea of cognitive networks of the brain [78]. This hypothesis predicts that mood congruence fosters cognitive processes based on the organization of our brain. A positive mood like enjoyment boosts the remembrance of Positive information (such as feedback after a successful test) than in a negative mood (like sadness); negative emotions on the other hand enhance the remembrance of negative information such as finding you have failed an exam. The architecture of our brain plays a major role in the power of congruence because the brain is organized by associations and semantic similarities: the closer is the location of the information and the easier the activation, the more similar and the stronger the association.

2. Schwarz (1990) The second theory suggested is the theory of “mood as information” which is based on the informative ability of a mood [79]. According to the theory, the important point is the variety of information that is embedded in the mood of a learner. Positive signifies the presence of positive characteristics in a situation while negative mood signals negative elements. Like a ‘How-do-I-feel?’ heuristic, a person interprets their mood and reacts positively in a positive mood and aversively in a negative mood.

3. The integration of the effects of the two theories above brought about the formation of the hypothesis of ‘mood-dependent cognitive styles’. It proposes that a positive mood indicates a pleasant and safe environment: a kind of environment that offers optimal preconditions for holistic and creative thinking as it enables open-mindedness.

To analyze the functions of emotions for school learning, a contextual shift from experimental research in laboratories to the classroom is necessary, as well as to connect the results from mood research to school research. What can be learned from mood research for school learning? The general message is that the fact that there are no simple effects of mood urges us to look into details. It encourages us not only to focus on learners' moods and learning outcomes but also to look at processes. The processes, in this case, are regarded as mediator variables, they tend to stimulate motivation in a learner. Figure 3.1 below presents some common mediator variables.

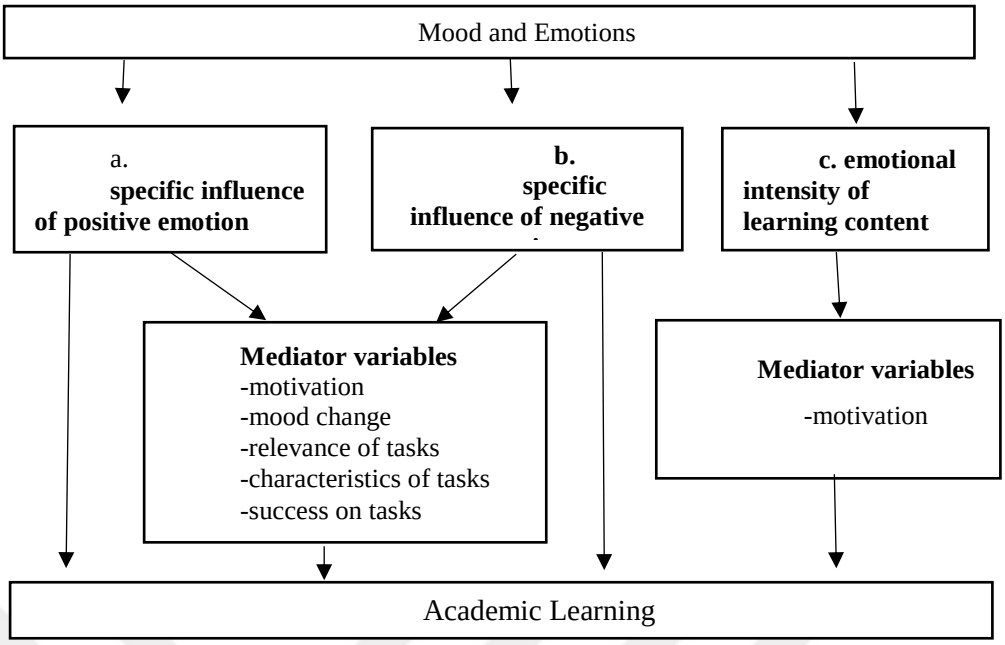


Figure 3.1. Effect of mood and emotion on school learning

The teacher's mood can evoke the learning mood in students during the process of learning. A negative impulse from a teacher might be more influential than a negative impulse from a learning topic. The concept of these academic emotions has been examined for almost two decades [16] and is presented in table 3.1. From the findings in this research, positive emotions foster all-inclusive, creative ways of reasoning. Harmful effects can only come to play when the students are in a state of that positive emotions but have less interest in the topic of discussion.

Table 3.1. Academic emotions

	Valance	
Activation	Positive	Negative
Activating	Enjoyment	Anxiety
	Hope	Anger
	Pride	Shame/Fault
Deactivating	Relief	Boredom
		Hopelessness

In their circumplex model of learning, Kort et al [80] have suggested a four-quadrant model that has six possible emotion axes: anxiety-confidence, ennui-fascination, frustration-euphoria, dispirited enthusiasm, terror-excitement, humiliated-proud: that may arise in the course of learning, relating phases of learning to emotions. This is depicted in the Figure 3.2 below. hence, "A typical learning phase includes a variety of emotions, cycling students around this four-quadrant cognitive-

emotive space as they learn. It is very useful to recognize that a range of emotions occurs naturally in a real learning process, and it is not simply the case that the positive emotions are the good ones. The negative half is an inevitable part of the learning cycle.

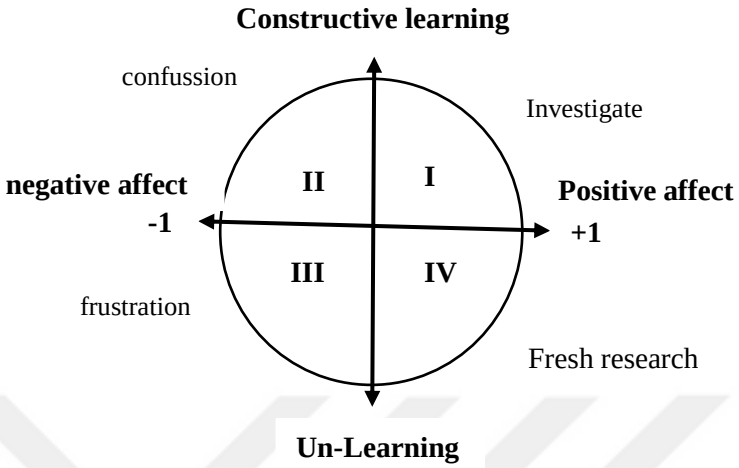


Figure 3.2. Learning Cycle model

3.3. Emotion Detection

There are two different approaches usually followed in emotion research: The information processing approach and the interactional approach [81]. This research follows the information processing approach that treats emotion as an entity similar to information that is communicated from one person to another. This is because, whenever we speak, our speech is directed towards a person or group of persons, and is always accompanied by a specific emotion. The listener perceives this emotion according to the level of its excitation. With regards to this approach, Scherer et al [82] has classified emotion research into three major schools:

1. The basic emotion (patterns are equivalent to basic emotions that are universally recognized e.g. fear, love, anger, happiness),
2. The emotion dimension (quantifies emotion using various dimensions, e.g., arousal, valence, intensity, etc.), and
3. The eclectic approach (use of verbal labels that seem appropriate to the aims of a particular study e.g. academic emotions).

Irrespective of the emotional school, emotion detection aims to retrieve an embedded class of emotion from a speech. In line with the above emotion theories, several tools have been developed and certain approaches adopted to capture the affective state of a user. In the sections that follow, is a review of different tools and methods used to capture the respondent's emotional state. In our analysis, we adopted the criteria in Feidakis [83]: based on the criteria, the tools found

in the literature to capture the user's emotion signals or affective state can be grouped into three areas: Bio-physiological, MotorBehavioural, and Self-Report:

- Self-Report: the use of a pictorial scale or verbal scales, and questionnaires for 1st person, subjective report
- Bio-physiological: Capturing biomedical signals such as electromyogram-EMG, electrodermal activity-EDA, electrocardiogram-EKG or ECG, electrooculogram-EOG, blood volume pulse-BVP, etc through the use of sensors.
- Motor-Behavioural: monitoring or recording of motor-behavioral activities such as voice intonation, facial expressions, body posture, sentiment analysis of text input, mouse and keyboard logs, etc.

There is not a substantial measurement tool that fully qualifies all these design requirements. Take self-reporting for instance, it is the only way to measure a user's subjective feelings. But users are often hesitant to divulge their inner feelings to researchers to avoid embarrassment [84]. Self-reporting is also considered to be intrusive, usually interrupting the learning process. The bio-physiological tools on the other hand are good means of collecting emotional information, but they are generally more expensive and time-consuming. The best tool for this research is the motor-behavioral tools to capture the motor-behavioral activities especially voice intonation and other voice properties needed. This way we analyze the voice to detect the embedded emotion. Another way is to carry out a sentiment analysis of text input, however is a very difficult natural language processing(NLP) task applying an information retrieval approach [85]. Another strategy of text analysis is the application of supervised learning and classification algorithms, such as support vector machines [86] or latent semantic analysis [87], to develop statistical models for identifying the emotional content of texts [88]. The setback with the supervised learning approach is its requirement for relatively large quantities of manually tagged samples [89].

The third type of strategy is based on the use of the affect dictionary that contains a dictionary of words with a reasonable portion of affect in the language that is being analyzed. These words may act as 'triggers' for expressions of emotion. The recovery of such expressions of emotion can be enhanced using lexico-semantic resources like WordNet [90] and multilingual linguistic resources like FreeLing [91]. These tools originate from Clinical Psychology and employ verbal and non-verbal descriptions of emotions. Usually, they are cost-free or inexpensive.

3.4. Speech Emotion Recognition

There are different ways through which humans communicate. However, the speech signal is one of the fastest, reliable, and most natural means through which humans communicate. Therefore the speech can be a fast and efficient method of interaction between humans and

machines. Speech emotion recognition (SER) is the study and analysis of speech signals via a collection of carefully selected methodologies to detect, process, and classify the embedded emotions in a speech signal. It is a branch of affective computing that aims to pass the human capability of emotional intelligence to machines in other to aid human-machine interaction (HCI). The expression of emotion has a significant role in complex social communication. When humans communicate, there is always one form of emotion or the other embedded in their speech. The emotions expressed during speech are most times expressed in the vocal characteristics and vocal interferences of the speech sound which are two classes of paralinguistic properties. Each of the characteristics that are found in the two paralinguistic categories carries an emotional cue that is explored during a speech emotion recognition task to detect the embedded emotion in the speech. SER is simply a technique that can recognize emotions in a speech. An abstract conceptual baseline architecture of a generic system that is capable of recognizing emotions in speech is composed of several processes and elements as presented in Figure 3.3. below. Speech emotion recognition is a pattern recognition task that just like many machine learning tasks compose of data processing, feature engineering, and classification.

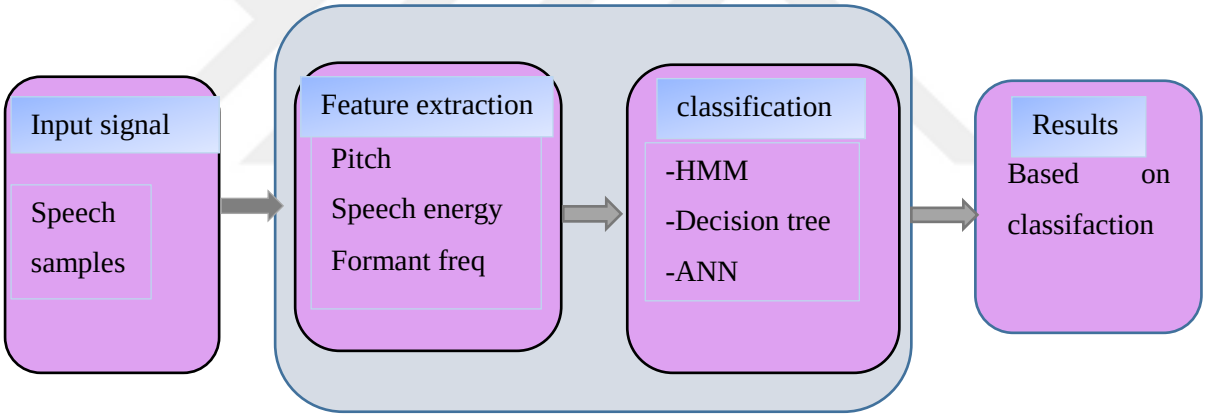


Figure 3.3. Speech emotion recognition framework

There are three steps in the general architecture of a Speech Emotion Recognition (SER) system. This architecture is shown in Figure 3.4.

a) A speech processing system that extracts some suitable features from the speech signal, such as pitch or energy, etc. This step is called the feature extraction or feature generation step.

b) The second step involves the summarization of these generated features into a reduced set of features with the help of a feature selector. It is called the feature selection step.

c) The third step utilizes a machine learning classifier that learns in a supervised manner with example data how to associate the features to the emotions. A deep-learning algorithm can also be used to detect the type of emotion associated with a feature. This step is called the classification step.

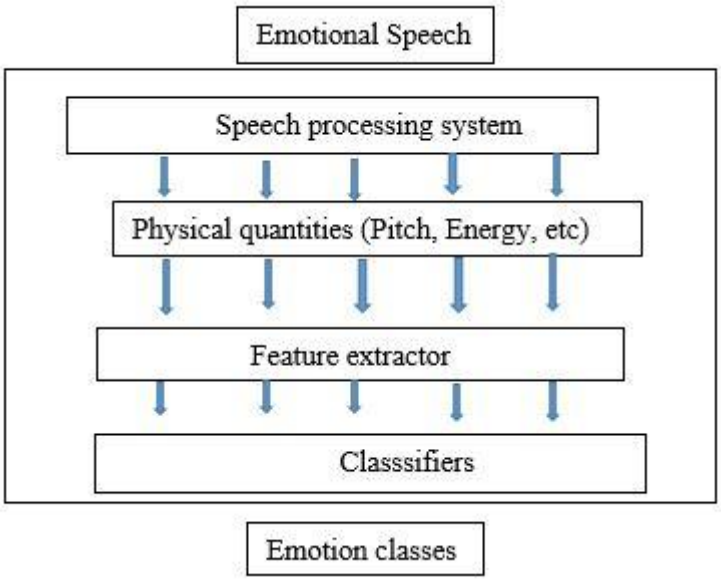


Figure 3.4. SER architecture

Both spectral and prosodic features are suitable for speech emotion recognition tasks because both of these features hold the affective information. The potential features are generated from each speech utterance for the computational mapping between emotions and speech patterns. The selected features are then used for training and testing by using any classifier method to recognize the emotions. The ER model of the training and validation of the training model is depicted in Figure 3.5.

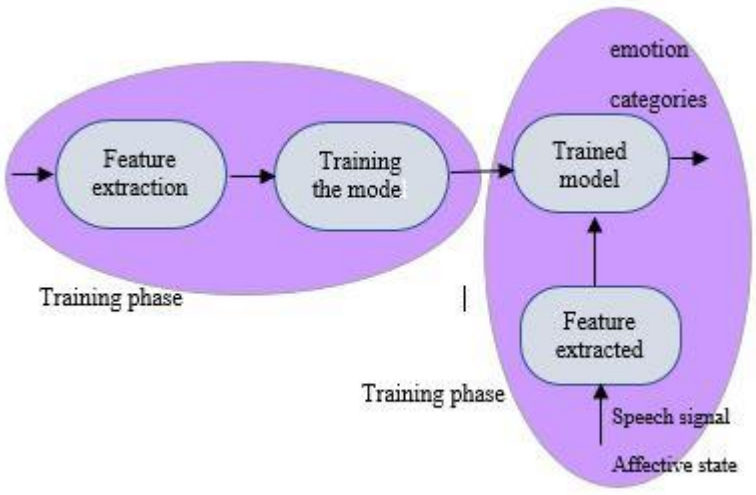


Figure 3.5. Training and validation of ER models

3.4.1. Application of Speech Emotion Recognition

The application of machine intelligence in the sub-domain of audio analysis is growing very rapidly. It improves the functionality of digital systems and machines. For example in virtual assistants such as Amazon's Alexa, Apple's Siri, and Google Home, are systems that perform critical artificial cognition tasks from audio data. Speech emotion recognition is also highly adopted in other fields as well. In customer service, for example, SER is integrated into the system to be able to recognize the affective state of the customers and adapt their responses accordingly [92-94]. It also finds application in robotics where it is used in the design of intelligent household robots [95]. Speech emotion recognition also plays an important role in medical psychology to aid the treatment of psychiatric patients. It also finds application in defense and security, especially in an investigation process where it aids the interrogation process after a crime is committed.

SER also finds intensive application in educational settings where it is integrated into the educational system to improve learning. In this case, it is utilized to build an interactive learning platform to make learning easier. The application is an educational setting, however, is mostly focused on recognizing the affective state of the learning or user to adjust the learning material that will enhance the learning ability of the learner. In this work, we focus on the application of speech emotion recognition in distance learning where we try to detect the emotional state of a teacher or a lecturer during the process of teaching.

The detection of a lecturer's emotion is also important because it plays a major role in enhancing the learning process for the students. Teachers are regarded as the most significant figure of any educational institution, and their enthusiasm is very important for students in the classroom. In line with the inquiries of teacher enthusiasm, principles of PP, and classroom enjoyment [3]. Teaching is an occupation that requires a reasonable level of emotional labor. It requires a lot of effort in planning and control. Teachers need to express organizationally desired emotions during the process of teaching. An unpleasant emotional classroom atmosphere can have considerable implications for students learning, school climate, and quality of education in general.

The method we used is generic and can be extended to emotion detection tasks in any kind of settings giving the proper datasets.

3.4.2. Data Acquisition for SER

There are many speech processing tasks and each of these tasks requires a suitable dataset for the effective functioning of its model. A speech emotion recognition is not an exception as it requires a dataset that comprises speech utterances that have different kinds of emotion embedded in them. Many existing datasets have been extensively used in speech emotion recognition tasks such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),

Interactive Emotional Dyadic Emotion Capture Database (IEMOCAP), Database of German Emotional Speech (EMO-DB), Acted Emotional Speech Dynamic Database (AESDD), etc. For this research, we will be collecting a new dataset that embeds the different kinds of emotions that are required for our work. The sub-sections below describe the emotions and the vocal cues that comprise these emotions in our bid to collect a very sound dataset. The dataset for this task should comprise of emotions that are relevant in academic settings, so they are called academic emotions.

3.4.3. Academic Emotion Dataset

Every speech processing application employs certain properties or features of speech signals in other to achieve its target goal. speech is part of a multichannel system involved in conveying emotion. Understanding how it operates in that context requires suitable data, consisting of multimodal records of emotion drawn from everyday life. It is assumed that there exist several quantifiable vocal parameters that reveal the affective state an individual is experiencing at a particular moment (or expressing for strategic purposes in social interaction). There is a large amount of truth in this assumption, given that most emotional states involve physiological reactions (e.g., changes in the autonomic and somatic nervous systems), which in turn modify various areas of the voice production process. For instance, the muscle tension is increased and a noticeable change in breathing occurs when aroused with anger. This influences the vibration of the vocal folds and vocal tract shape, affecting the acoustic characteristics of the speech, which in turn can be used by the listener to infer the respective state [96]. Speech contains nonverbal elements which are referred to as paralanguage e.g Voice quality, Speech rate, pitch, volume as well as other prosodic features such as rhythm, intonation, and stress. These elements are used to modify meanings and often convey relational information such as feelings and emotions. Humans can easily understand these emotions during communication. The recognition rate improves with an increase in age. From research[97], it was found out that children overall can identify the emotional sounds with an average accuracy of 78.1% for the younger children and 83.9% for the older children [97]. This level of performance is much higher than would be expected by chance in a four-way forced-choice task (approximately 25%)

To test our proposed model of ShoePatNet23, two SER datasets were collected during a lecture process. The difference between these datasets is the languages used. In the first and second speech datasets, Turkish and English languages were used respectively.

In other to obtain a dataset that is suitable for this particular task of lecturer's emotion recognition in a distance learning education system, we decided to collect our dataset. To collect this dataset, we consider another class of emotional labels that have a direct correlation with a typical interactive learning atmosphere. During a learning session, the lecture could be boring, interesting, and neutral. These states are often influenced by the level of expressiveness of the

teacher. How clear is the message, what is the excitation level of his speech? these and many others are greatly influenced by the lecturer's emotional state. Furthermore, because the different classes of emotions could be grouped into negative, positive, and neutral, we combine this knowledge to come up with a new label of educational or academic emotions: Positive, negative, and neutral. Each of these affective states creates a different kind of learning atmosphere and therefore influences learning differently. So, in collecting the dataset we consider the different characteristics of the presented emotional states. What features of speech constitute positive, negative, and neutral emotional states. To do these, we consider the paralinguistic features (especially the vocal characteristics) such as and the prosodic features such as rhythm, intonation, and stress. Experience the emotional labels.

3.4.4. Voice Emotional Features

The expression of emotion in the voice can be analyzed at three different levels: The physiological level (e.g., describing nerve impulses or muscle innervation patterns of the major structures involved in the voice-production process), the phonatory-articulatory level (e.g., describing the position or movement of the major structures such as the vocal folds), and the acoustic level (e.g., describing characteristics of the speech waveform emanating from the mouth).

At the moment, the majority of the current measurement methods that are employed at the physiological and phonatory-articulatory levels are rather obtrusive and need specialized equipment as well as a high level of know-how. In contrast, acoustic cues of vocal emotion expression can be acquired objectively, economically, and discreetly from recorded speech, and permit some suggestions about voice production and physiological elements. Hence, acoustic measurement of voice cues is perhaps the best method since it holds the greatest promise for interdisciplinary research on emotional speech, even though it requires basic training in voice physiology and acoustics but still does not need special equipment.

In the study of vocal cues, Voice cues are usually divided into those related to:

(a) Fundamental frequency (F0, relates to the perceived pitch),

(b) Vocal perturbation (short-term variability in sound production),

(c) Voice quality (a correlate of the perceived 'timbre'),

(d) Intensity (a correlate of the perceived loudness), and

(e) Temporal aspects of speech (e.g., speech rate), as well as various combinations of these aspects (e.g., prosodic features).

It is beneficial to determine the different voice descriptors that exist in different emotions that can be used to infer from the voice the affective state of the speaker. Voice production is however complex and poses a challenge to the effective detection of these voice profiles. Several factors lead to these complications that mare the search for voice cues, for example, interactions

between unprompted and strategic expression, important variations within particular emotion families (e.g., hot vs. cold anger), individual differences among speakers, and also the verbal content. Unsurprisingly, many studies commonly highlight the inconsistent data regarding vocal cues to specific emotions. Some studies had gone ahead to propose that arousal is the only vocal cue embedded in a voice because overall arousal (such as high fundamental frequency F0 and fast tempo) is the most agreed upon voice element. There is, however, considerable proof that voice cues can differentiate affective states beyond the simple aroused/sleepy (arousal), and pleasant/unpleasant (valence) affective dimensions of activation.

Many works in literature employed emotions displayed by professional actors, but the important question is to what extent does such portrayals deviate from the natural expressions. This question is yet to be completely answered because just a few researchers have attempted to make a direct comparison between the two types of speech samples. Furthermore, most studies always focus only on a few basic emotions, while neglecting more complex emotions. Hence, much of the pertinent work on emotion differentiation in the voice remains to be done [98].

There is a collective understanding regarding the relevance of evolutionary methods to an understanding of emotional speech, but most of the works done so far are theoretical in approach [99]. Many proposed affect theories do not make explicit and detailed predictions regarding the specificity of the affective states embedded in an emotional speech, but they can be expected to differ in terms of whether only dimensions like valence and arousal or distinct primary emotion classes are presumed to be vocally differentiated. Component appraisal theories predict that emotional speech will convey also finer nuances that reflect the precise cognitive appraisals and consequent action tendencies that underlie each emotion. One emotion theory that has been widely accepted so far and which has received preliminary support in some studies is the work of Scherer's in 1986 [100] component process theory[101]

Emotion in speech is a communication system that has several parts:

1. The encoding (which is the expression or portrayal of the emotion by the speakers)
2. the acoustic cues (e.g., sound intensity) that convey the embedded or proposed emotion,
3. The perceived loudness (which is the proximal perception of the cues by the listener)
4. The decoding (which is the inference about the expressed emotion by the listener)

A comprehensive understanding of these parts is very important for a thorough understanding of speech emotion, although it is okay to research any part, per Brunswik's lens model [98].

Several studies have explored emotion cues from voice cues in listening test experiments. The participants in the experiments were asked to guess the emotions embedded in samples of

speech using response formats such as forced-choice, quantitative ratings, and free labeling. To determine a specific voice cue used in the expression of a particular emotion, different content-masking procedures like low-pass filtering were employed [99]. The sole purpose of the content-masking methods is to disrupt or degrade individuals during the analysis of the voice signal. Juslin et al [102] in the year 2003 presented the most extensive review to date on vocal expression. In their work, they studied 39 vocal expressions, featuring 60 listening experiments were included in a meta-analysis of decoding accuracy based on forced-choice judgments. Within-cultural and cross-cultural studies were included in the meta-analysis, and both portrayed and natural expressions. From the results of the analysis, the overall decoding accuracy for within-cultural expression was equal to a score of 70% correct in a forced-choice task with five feedback alternatives. Decoding accuracy was 7% higher for within-cultural than for cross-cultural expressions.

The result from the studies in the preceding section suggests that the type of speech sample used plays a major role in the estimates of decoding accuracy. Three different samples of voice have been used in previous research:

1. Emotion is displayed by professional actors in a controlled environment such as a laboratory or studios, encouraging experimental control and ensuring strong effects on voice cues but raising doubts about ecological validity.

2. Real-time capturing of vocal expressions in a natural environment or from reality media broadcasts. Ecological validity can be expected to be high (at least for unobtrusive recordings) but it is difficult to determine what emotional state is felt or portrayed by the speaker (often inferred from situational cues).

1. The third category is the experimentally induced emotion expressions in the laboratory which combines experimental control with the possibility of obtaining spontaneous emotion expressions. The induced affective states are often weak and unspecific.

The following are beneficial for obtaining emotion differences in a speech study:

a. analysis of a large number of voice cues,

b. precision in the labeling of the affective states expressed, and

c. a proper research design based on explicit predictions.

In particular, it seems important to go beyond single measures of the most common voice cues (e.g., F0, speech rate, intensity), which may involve similar cue levels for different emotions. Furthermore, it appears necessary to control for the emotional intensity, which may affect voice cues in a differential manner.

The details of these speech corpora are given in the next section

4. MATERIAL AND METHOD

4.1. Dataset

The dataset used for this task is collected for speech emotion recognition in a distance learning environment for the detection of lecturers' emotions. The dataset comprises three emotions which are positive, negative, and neutral which correspond to those emotions that make a lecture either interesting, boring, neutral respectively. In collecting this dataset, the participants were ensured to induce the different classes of emotions in their utterances using the understanding of voice emotional cues from the preceding sections to record interesting, boring, and neutral lectures. The emotional cues that were used mainly for the collection of the dataset are the Speech Rate, Pitch, and Tone of the voice. The explanations of the different classes are given below.

Interesting Lecture: An interesting lecture is usually characterized by positive emotions which have a high excitation level. The interesting speech comprises utterances with a high speech rate, high and lively tone, and a high pitch.

Boring lecture: A boring lecture is usually characterized by negative emotions with low excitation levels such as sadness, anxiety, and boredom. It comprises speech utterances with a very low speech rate, low pitch, and low tone of voice.

Neutral Lecture: A neutral lecture is somewhere between interesting and neutral. The lesson is neither interesting nor boring.

The dataset is collected in two different languages. The explanation of each of the corpus is given below.

4.1.1. Turkish Speech Dataset

In this work, a new SER dataset is collected and this dataset consists of 7101 sounds with a length of five seconds. The sounds are collected from variable mobile phones and these mobile phones have a single channel. These speeches are collected from 18 lecturers (8 females and 10 males). The collected speeches were stored as acc, m4a, mp3, mp4, Ogg, and AMR formats. There are three categories and in this dataset and these categories are named negative (-1), positive (+1), and neutral (0). We used a fixed lecture note for collected this dataset and this lecture is a Turkish digital forensic lecture. More than 10 minutes of speech recording for each emotion was taken from each lecturer. The collected speeches were divided into five-second pieces and these pieces consist of our observations. In this dataset, there are 2473 negative (-1), 2231 neutral (0) and 2397 positive (+1) speech observations. The distribution of the collected dataset per the categories is shown in Figure 4. 1.

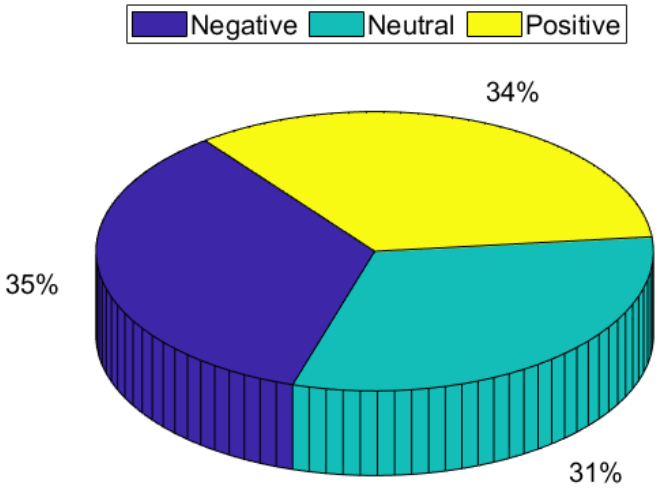


Figure 4.1. Distribution of the collected data

4.1.2. English Speech Dataset

The dataset was collected from a total of 45 speakers which comprises 15 females and 30 males. The speakers are within the age range of 24 to 42 years. The audio file has an average length per sound of eight minutes. The dataset was collected from these speakers while they were giving a lecture from a prepared lecture slide on digital forensics. Each speaker was made to give the lecture in three different affective states of interesting, boring, and neutral which forms the three classes to be used in classifying a lecture delivered via a distance learning system. The audio files were recorded using a mobile phone and were collected in five different formats of MP4, MPEG, AAC, OGG, and M4A. The interesting class has a total of 44 sounds and the neutral class has a total of 47 sounds while the boring has a total of 43 sounds. The table below gives the summary of the dataset.

Table 4.1. Dataset Summary

SOUND FORMAT	MP4	MPEG	AAC	OGG	M4A
No. of Speakers	45				
No. of Female Speakers	15				
No. of Male Speakers	30				

Average length per sound	8 minutes
No. of Class Labels	3: Interesting, Neutral, and Boring
Number of observations with a length of 5 seconds	Interesting/positive: 3446 Boring/negative: 2865 Neutral: 3230 Total: 9541

Table 4.1 above provides a detail summary or explanation of the collected dataset for the experiment.

4.2. The Proposed Method

A new effective speech emotion recognition model is presented in this research. This model is implemented using a shoelace graph pattern as a baseline feature generation network, a TQWT, a Chi2 selector, an NCA selector, and a Cubic SVM classifier. Therefore, the model is named ShoePatNet23. ShoepatNet23 is an effective model that consists of many concepts and is very successful for any speech emotion recognition task. The shoelace pattern is newly introduced, it is a one-dimensional local feature generation network. This feature generator is proposed to investigate the feature extraction ability of the different patterns used in knotting shoelaces using graph theory. To improve the capability of this feature extraction network, multiple kernels have been used to generate binary features comprehensively. A signal decomposition technique of tunable Q wavelet transform (TQWT) was employed on the speech signals in two different modes: high oscillatory mode and low oscillatory mode of decomposition. Deploying this two-phased TQWT technique generates 22 signal sub-bands, and this improves the number of features to be extracted. Therefore, the shoelace pattern feature generation network creates 22 different feature vectors from the generated signal sub-bands. A 23rd feature vector is generated from the raw speech signal, hence, the name ShoePatNet23. Each of the feature vectors has a length of 3072. In other to get the best features from the 3072 features of an individual vector, a Chi2 selector was applied which reduces the feature-length from 3072 to 512. To further improve the ability of the network by ensuring that we get the best informative features, the loss value for each feature vector is calculated after applying an SVM classifier to test the effectiveness of each of the 23 feature vectors. The top five feature vectors were chosen, and the features from these five vectors were merged. A neighborhood component analysis (NCA) method is further applied to the merged features to reduce their dimension to 512 features. Therefore the Shoelace feature extraction

network extracted the best 512 features (final features). These final features are fed unto the classifier algorithms we were able to obtain the highest classification rate using Cubic SVM. The applied ShoePatNet23 made a reasonable contribution in the following areas;

- It investigates four widely used shoelace patterns for local feature generation,
- Generation of signal sub-bands by the intelligent application of two-phased Tunable Q wavelet transform (2-P TQWT),
- An intelligent way of generating and choosing the most valuable feature vectors both in the feature generation phase and feature selection phase,
- Showing universal success of the ShoePatNet23 using two datasets with two languages.

Figure 4.2 gives the graphical illustration of the proposed ShoePatNet23.

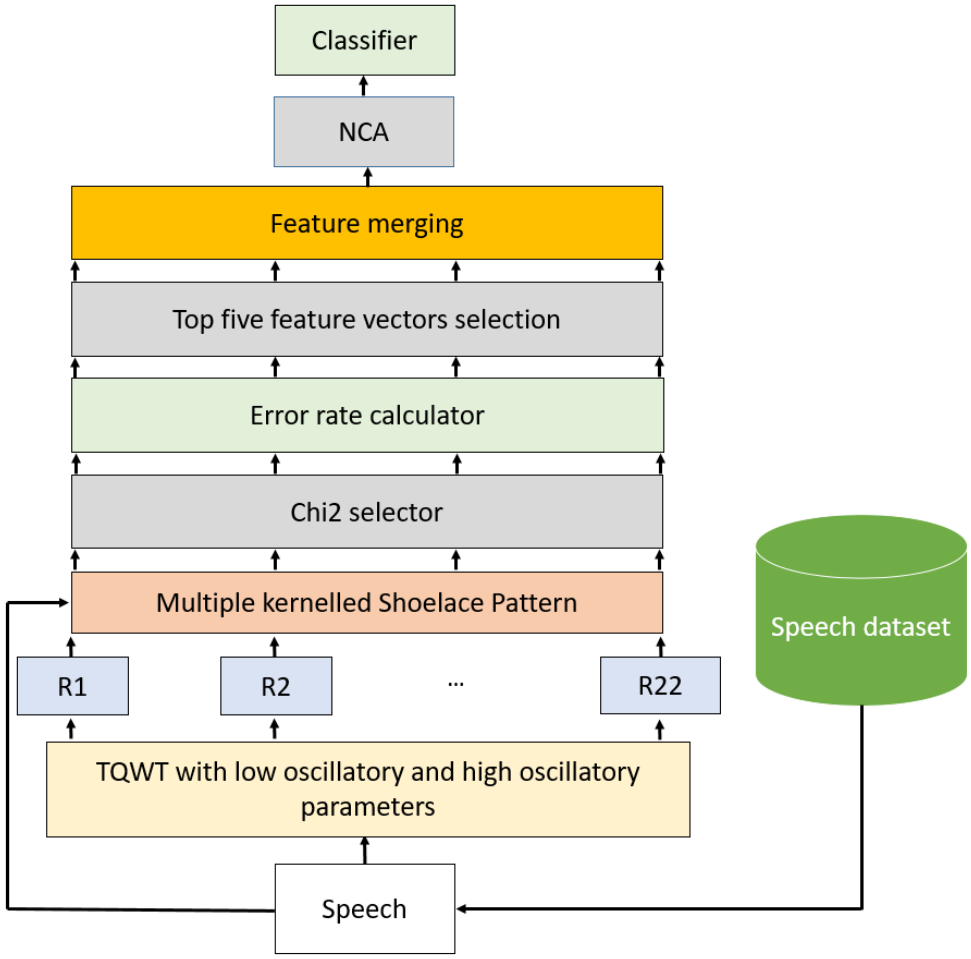


Figure 4.2. Schematic representation of the proposed ShoePatNet23

This model employed a tunable Q wavelet transform in high and low oscillation phases to generate sub-bands (R), thereby generating 22 sub-bands. The proposed shoelace feature extraction network is equipped with multiple kernels that aids with effective feature extraction. This network

is used to generate features from the 22 wavelet sub-bands and also from the original speech signal. It extracts a total of 3072 features from each feature sub-band. The Chi2 feature selector selects 512 out of the 3072 features and these feature vectors are tested to confirm the discriminative ability using a support vector machine (SVM) to calculate their loss values. The top five performing feature vectors are selected using the calculated loss values (misclassification rates) and these feature vectors are merged to obtain the final feature vector with a length of 512 × 5=2560. By deploying another feature section algorithm of NCA, the most valuable/informative 512 features are selected from the created 2560 features and they (the selected 512 features) are utilized as input of the classifier. To better explain our proposal, pseudo-code of our proposed ShoePatNet23 are given below.

Table 4.2. Algorithm for ShoePatNet23

Input: *The used speech datasets* ***Output:*** *Validation predictions (results)*
00: Load one of the used dataset. *01: Read each speech.* *02: Divide the speech into segments with a length of five seconds and obtain observation* *03: Apply TQWT to generate sub-bands.* *04: Extract 3072 features from the observation and create the first feature vector by applying the proposed multiple kernel shoelace pattern.* *05: Apply the proposed multiple kernel shoelace pattern to the created 22 sub-bands and create other 22 feature vectors.* *06: Apply Chi2 to the generated 23 feature vectors and choose the most informative 512 features of each feature vector.* *07: Calculate the misclassification rate of each feature vector and create a loss array with a size of 23.* *08: Select the top five feature vectors using the loss array.* *09: Concatenate the selected features and obtain 2560 features.* *10: Select the top 512 features from the generated 2560 features by applying NCA.* *11: Classify the selected 512 features applying Cubic SVM.*

Table 4.2 above describes the procedure for the entire processes involed in the proposed model, from speech processing to classification. More details of the proposed ShoePatNet23 are given in below.

4.2.1. Feature Extraction

The most important phase of the proposed Model (ShoePatNet23) is the ShoePatNet23 is feature generation phase because the distinctive nature of the feature extracted determines the quality of the prediction outputs. We developed an effective machine learning technique that will extract the most discriminative feature vectors. Here in this method, the popular graph theory is integrated into the network and utilized to propose a new feature generator network. The graph pattern adopted model four shoelacing concepts, i.e it is designed in four different ways through which we knot the shoelace on our shoes in a graphical manner. This research, therefore, shoes the feature extraction ability using graphs, in particular, shoelace graphs. However, the newly introduced shoelace pattern is a hand-crafted local feature extractor. It is important to enhance the feature generation ability of the shoelace pattern to improve the classification ability of our entire model, therefore, tunable Q wavelet transform is firstly applied on the speech signal to decompose it to yield higher-level feature vectors. TQWT though has a certain limitation due to its parameter tuning problem. In other to take care of this defect, the concept of 2-phased TQWT is introduced where it is applied in high oscillatory and low oscillatory phases. This decomposes the signal into 22 sub-bands. The signal components that are generated from these phases are then used to obtain the feature vectors. The 22 sub-band signals alongside the raw speech signals combined to give a total of 23 feature vectors. Shoelace pattern network is applied on the 22 sub-band signal components as well as on the raw speech signals to extract features for the network. This is why the machine learning model is named ShoePatNet23. Upon the application of the shoelace graph patterns, a total of 3072 features are generated with each feature vector. To reduce the dimension of these features generated at this stage, a Chi2 selector is employed on the features and the length of whole feature vectors is reduced from 3072 to 512. It is always beneficial to limit the number of features to the bearest minimum containing the most informative features, therefore, at every step of feature generation, we try to discard those features that don't carry much information about the signal. We did the same with the 23 feature vectors to determine which feature vectors yielded more discriminative features by computing the misclassification rates (error value) of all the after testing with a support vector machine (SVM) classifier. We might have utilized the SVM as the misclassification rate calculator which is a parametric method, other misclassification rate calculators also exist that people can try out. Judging from the results obtained especially the losses as indicated by the misclassification ratios, the top five performing feature vectors were selected and the features from these vectors are concatenated to give a total of 2560 features. The following describes the implementation steps of the feature extraction network.

Step 0: Load speech signal.

Step 1: Create segments with a length of five seconds.

$$sp = s(i + j - 1), i \in \{1,5 \times fs, \dots, len(s) - 5 \times fs + 1\}, j \in \{1,2, \dots ,5 \times fs\} \quad (4.1)$$

where sp is created speech segment, s is speech signal, fs defines the frequency of the speech signal, $len(s)$ is the length of the speech, i and j are indices.

Step 2: Apply the tunable Q wavelet transform (TQWT) to each speech segment. TQWT is an effective signal decomposition method with tunable wavelet parameters. The parameters used here are the three TQWT parameters: oscillatory value (Q), redundancy (r), and the number of levels (J). We used two different groups of parameters to generate low oscillatory and high oscillatory sub-bands. these parameters are given as follows.

Q1=1, r1=3, J1=8, Q

2=4, r2=3, J2=12.

Applying these sets of parameters generated 22 sub-bands from the network. Each signal parameter generates 11 sub-bands

$$R1 = TQWT(sp, Q1, r1, J1) \quad (4.2)$$

$$R2 = TQWT(sp, Q2, r2, J2) \quad (4.3)$$

$$R = merge(R1, R2) \quad (4.4)$$

Herein, $TQWT()$ denotes tunable Q wavelet transform decomposition function. It takes four parameters. These parameters are the used signal, Q, r, and J respectively. $R1$ is the generated low oscillatory sub-bands and $R2$ represents high oscillatory sub-bands. By applying the merge function ($merge$), a data structure is created and this data structure has 22 sub-bands ($R1$ has 9 sub-bands and $R2$ has 12 sub-bands).

Step 3: Generate features from raw speech signal sp and each R deploying multi-kernel shoelace pattern.

$$fv^1 = ShoePat(sp) \quad (4.5)$$

$$fv^{k+1} = ShoePat(R^k), k \in \{1,2, \dots ,22\} \quad (4.6)$$

In the equation above, fv^k is the generated k[th] feature vector. Deploying the shoelace pattern ($ShoePat$) generates features with a length of 3072. The section below gives a detailed explanation of the proposed shoelace pattern feature extraction function and the tunable Q wavelet transform (TQWT) signal decomposition function

ShoePatNet

The shoelace pattern is modeled based on graph theory which is a popular mathematical theory that has many applications in real-life situations. Over the years, mathematicians have tried to understand the concept of shoelace patterns and their effective applications in mathematics. It proved difficult until recently when authors put forward a solution for this problem [103]. Following this breakthrough, we decided to adopt these patterns to generate textural features from the speech signal. The proposed feature extractor employed four different patterns by which shoelaces are knotted (see Figure 3) as pattern, therefore, this feature extraction function is named

shoelace pattern. In other obtain the feature matrix, eight-sized overlapping blocks were applied to obtain a 4 x 2 sized matrices feature generation pattern. The *eight*-sized overlapping block and the 4 x 2 sized matrix are used to develop the shoelace pattern as demonstrated in Figure 4.3.

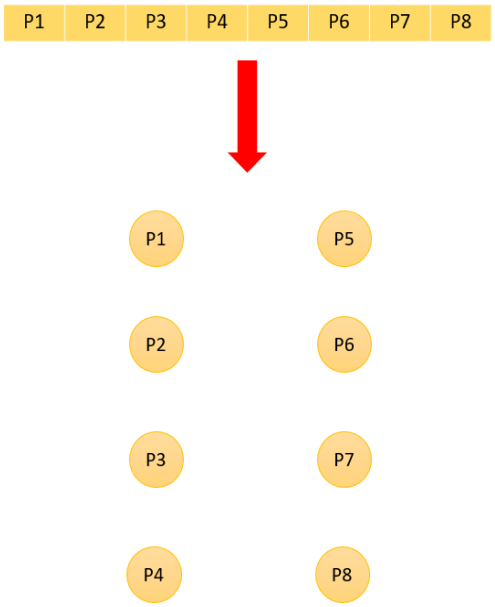
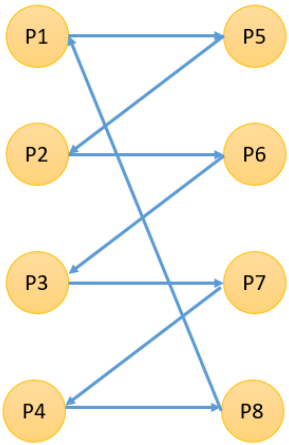
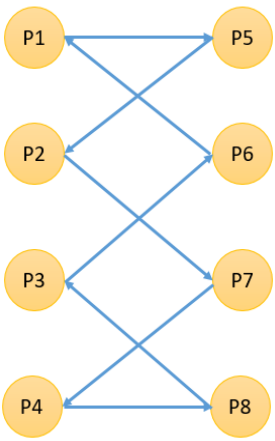


Figure 4.3. Overlapping block of 4 x 2 sized matrix for shoelace pattern

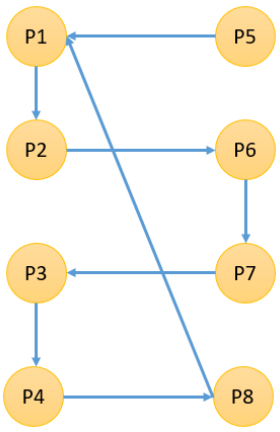
Judging from Figure 2, the point values of the overlapping window are considered as the node of a graph. Using the node concept of a graph alongside the shoelace patterns, four graphs were generated to represent four commonly used shoelace knotting patterns. These four graphs, therefore, are the baseline concept in the development of the newly presented model. The shoelace patterns that were used are presented in Figure 4.4. Below.



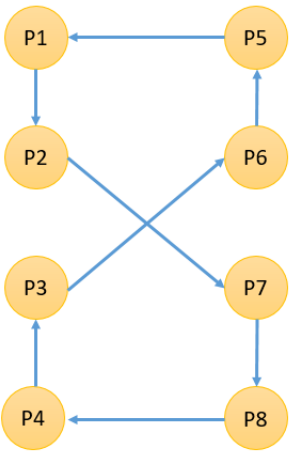
(a) First shoelace pattern



(b) Second shoelace pattern



(c) Third shoelace pattern



(d) Fourth shoelace pattern

Figure 4.4. Illustration of Shoelace patterns used in ShoePatNet23

These kinds of generated graphs are referred to as directed Hamiltonian graphs. The starting node of the first and second graphs (shoelace patterns) is P1, while the starting node of the third and fourth shoelace pattern is P5.

As can be seen from the graphs above, eight edges are present in each graph. Each of the edges represents an input value to the signum function. hence, each pattern generates *eight* bits using each binary feature extraction/generation function. For this research, we used both ternary and signum kernels to generate binary features from each pattern. In application, the ternary kernel has two binary feature extraction functions that are named the upper ternary function and the lower ternary function. Therefore, in the end, we have three binary feature generation functions. The mathematical representation of these three binary feature extraction functions is given in Equations (7) - (10).

$$\zeta^1(t,r) = \begin{cases} 0, t-r < 0 \\ 1, t-r \geq 0 \end{cases} \tag{4.7}$$

$$\zeta^2(t,r) = \begin{cases} 0, t-r \leq d \\ 1, t-r > d \end{cases} \tag{4.8}$$

$$\zeta^3(t,r) = \begin{cases} 0, t-r \geq -d \\ 1, t-r < -d \end{cases} \tag{4.9}$$

$$d = \frac{Std(sp)}{2} \tag{4.10}$$

Where $\zeta^1(.,.), \zeta^2(.,.)$ and $\zeta^3(.,.)$ are signum, upper ternary, and lower ternary functions respectively. These functions are applied as the kernel of the shoelace pattern. t, r are input values and $Std(.)$ represents standard deviation calculation function.

By employing the shoelace patterns presented in figure 3 above along with the defined kernels (see Figure 3 and equations (7) – (10)), we generated twelve map signals. The implementation steps of the proposed shoelace pattern are given below.

Step 3.1: Apply the kernel functions defined above to generate bits

$$\begin{bmatrix} bit_1^k \\ bit_2^k \\ bit_3^k \\ bit_4^k \\ bit_5^k \\ bit_6^k \\ bit_7^k \\ bit_8^k \end{bmatrix} = \zeta^k \left(\begin{bmatrix} P1 & P5 \\ P5 & P2 \\ P2 & P6 \\ P6 & P3 \\ P3 & P7 \\ P7 & P4 \\ P4 & P8 \\ P8 & P1 \end{bmatrix} \right), k \in \{1,2,3\} \tag{4.11}$$

$$\begin{bmatrix} bit_1^{k+3} \\ bit_2^{k+3} \\ bit_3^{k+3} \\ bit_4^{k+3} \\ bit_5^{k+3} \\ bit_6^{k+3} \\ bit_7^{k+3} \\ bit_8^{k+3} \end{bmatrix} = \zeta^k \left(\begin{bmatrix} P1 & P5 \\ P5 & P2 \\ P2 & P7 \\ P7 & P4 \\ P4 & P8 \\ P8 & P3 \\ P3 & P6 \\ P6 & P1 \end{bmatrix} \right) \tag{4.12}$$

$$\begin{bmatrix} bit_1^{k+6} \\ bit_2^{k+6} \\ bit_3^{k+6} \\ bit_4^{k+6} \\ bit_5^{k+6} \\ bit_6^{k+6} \\ bit_7^{k+6} \\ bit_8^{k+6} \end{bmatrix} = \zeta^k \left(\begin{bmatrix} P5 & P1 \\ P1 & P2 \\ P2 & P6 \\ P6 & P7 \\ P7 & P3 \\ P3 & P4 \\ P4 & P8 \\ P8 & P1 \end{bmatrix} \right) \tag{4.13}$$

$$\begin{bmatrix} bit_1^{k+9} \\ bit_2^{k+9} \\ bit_3^{k+9} \\ bit_4^{k+9} \\ bit_5^{k+9} \\ bit_6^{k+9} \\ bit_7^{k+9} \\ bit_8^{k+9} \end{bmatrix} = \zeta^k \left(\begin{bmatrix} P5 & P1 \\ P1 & P2 \\ P2 & P7 \\ P7 & P8 \\ P8 & P4 \\ P4 & P3 \\ P3 & P6 \\ P6 & P5 \end{bmatrix} \right) \tag{4.14}$$

Step 3.2: carry out a binary to decimal conversion by calculating map signals values.

$$map_t^j = \sum_{n=1}^{8} bit_n^j * 2^{n-1}, j \in \{1,2,\dots,12\}, t \in \{1,2,\dots,Ln-7\} \tag{4.15}$$

where map_t^j is t[th] value of the j[th] map signal and Ln is the length of the speech segment.

Step 3.3: Generate the histograms ($Hist^j$) of every map signal.

$$Hist^j(f) = 0, f \in \{1,2,\dots,2^8\} \quad (4.16)$$

$$Hist^j\left(map_t^j\right) = Hist^j\left(map_t^j\right) + 1 \quad (4.17)$$

Step 3.4: Concatenate the histograms from the step above to generate shoelace pattern features ($feat$) with a size of $256 \text{ x } 12 = 3072$.

$$feat = Hist^1 \mid Hist^2 \mid \dots \mid Hist^{12} \quad (4.18)$$

The defined *four* steps (Steps 3.1-3.4) are the steps that create the $ShoePat(.)$ feature generation function, and the $ShoePat(.)$ extracts 3072 features from a one-dimensional signal.

Step 4: Apply a Chi2 feature selector to Select 512 out of the 3072 features of each feature vector. Chi2 feature selector is a very fast feature selection function in the literature. Therefore, we employed the Chi2 selector.

$$ind^j = fselChi2\left(fv^j, y\right), j \in \{1,2,\dots,23\} \quad (4.19)$$

$$rf^j(:,t) = fv^j\left(:, ind^j(t)\right), t \in \{1,2,\dots,512\} \quad (4.20)$$

Herein, ind^j are qualified indexes of the fv^j, y represents real outputs/labels and rf^j is the j[th] reduced feature vector.

Step 5: Use a support vector machine classifier with 5-fold cross-validation to compute the misclassification rates of the reduced feature vector.

$$mr(j) = SVM\left(rf^j, y\right) \quad (4.21)$$

where mr is an error rate array with a length of 23.

Step 6: Select the top five performing reduced feature vectors (brf) using the result from the miscalculation rate (mr.) This step is parametric.

Step 7: merge brf to obtain a feature vector.

$$gf\left(l + 512(p-1)\right) = brf^p(l), p \in \{1,2,\dots,5\}, l \in \{1,2,\dots,512\} \quad (4.22)$$

where gf is the generated feature vector with a length of 2560.

Tunable Q wavelet transforms: Tunable Q-factor wavelet transforms (TQWT) is a signal decomposition technique. It is designed for analyzing oscillatory signals and uses flexible and fully discrete wavelet transform (DWT)[104]. TQWT is a flexible wavelet transform technique because of its adjustable input parameters: the Q-factor (Q), the redundancy or over-sampling rate (r), and the number of levels of the decomposition (J). The flexibility in wavelet function is achievable due to these tunable parameters. The TQWT and rational-dilation wavelet transform (RADWT) are very similar[105], the difference is that the TQWT doesn't require that the dilation factor be rational. The Q-factor, represented as Q, determines the oscillatory ability of the wavelet; specifically, Q determines the degree to which the oscillations of the wavelet are persistent. Q may be said to be a measure of the number of oscillations the wavelet displays. The acceptable value for Q is from 1.0 above (a real-value). The Q-factor of an oscillatory pulse is expressed as the ratio of

its center frequency to its bandwidth,

$$Q = \frac{f_0}{BW} \tag{4.23}$$

The redundancy of the TQWT is denoted as r and is computed using an infinitely high number of levels. The redundancy refers to the total over-sampling rate of the transform. This is computed as the ratio of total wavelet coefficients to the length of the signal to which the TQWT is applied. The desirable value of r is always greater than 1.0 and more desirable to go with values of 3.0 above. With a value of r around 1.0, the wavelet will not be well localized in time and is considered undesirable, because at these values it has too much ringing.

The number of stages (or levels) of the wavelet transform is denoted by J. The transform consists of a sequence of two-channel filter banks that when applied to a signal (sp) decomposes the signal into low-pass and high-pass sub-bands at every stage, with the low-pass output signal of each filter bank being employed as the input to the next filter bank. J is the number of filter banks. Each output signal represents one sub-band of the wavelet transform. Given a J level of decomposition (J filters), J + 1 sub-bands will be created iteratively by the two-channel filter bank. A 3-stage wavelet transform is illustrated in the Figure 4.5 below.

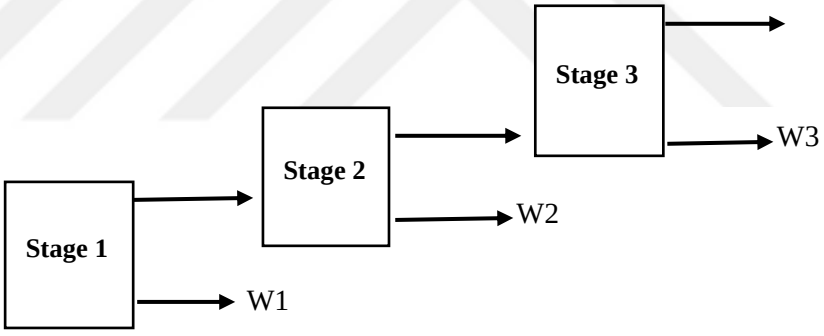


Figure 4.5. Phases of TQWT

The tunable-Q wavelet transform (TQWT) is implemented with the multi-rate filter bank illustrated in fig below. At every stage, the filter decomposed the signals into low-pass sub-band signal $v_0(n)$ and high-pass sub-band signal $v_1(n)$. Each of the sub-band has its sampling rate, low-pass sub-band has αf_s and high-pass sub-band has βf_s. Where f_s is the sampling rate of the input signal. The $\alpha \; and \; \beta$ are scaling parameters that must satisfy the condition:

$$0 < \alpha < 1 \quad and \;\; 0 < \beta \leq 1 \tag{4.24}$$

The selection of different parameters in TQWT determines its performance in getting the most informative features from the signal. The TQWT characteristic equation can be expressed as follows:

Chi2 Selector: It is a library package that computes the relationship between the target and the features. the method only selects the desired number of the variable with the best chi-squared values.

4.2.2. Feature Selection

Also known as variable selection, attribute selection, or subset selection, feature selection is a concept in machine learning that is used to choose specific variables or data points to maximize the efficiency and quality of prediction of a machine learning algorithm in this type of advanced data science. At the feature selection phase of the model, redundant or irrelevant features are discarded. This culling can make the output of our machine learning results stronger by maximizing the prediction accuracy. There are different feature selection tools and algorithms that help in creating very accurate models to improve machine learning outcomes, we, however, choose to use the neighborhood component analysis (NCA) algorithm for our proposed model. The NCA is one of the commonly implemented feature selectors in the literature and is a selection version of the nearest neighbor (1-NN) algorithm. It is a non-parametric method for selecting features to maximize prediction accuracy. For a multi-class classification problem with a training set that contains *n* observations as below:

$$S = \{(x_i, y_i), i = 1,2, \ldots\ldots n\} \tag{4.25}$$

where $xi \in \mathbb{R}p$ is the feature vectors, $yi \in \{1,2,\ldots,c\}$ are the class labels, and *c* is the number of classes. The purpose is to develop an algorithm $f: \mathbb{R}p \rightarrow \{1,2,\ldots,c\}$ which receives a feature vector and produces a prediction $f(x)$ for the true label *y* of *x*.

Consider a randomized classifier that:

- Picks a point *S* at a random as a reference point for *x* Ref(*x*).
- Using the label of the reference point Ref(*x*) labels *x*.

This technique is analogous to the implementation step of the K-NN classifier with a K value of 1. The nearest neighbor of the new point *x* is taken to be the reference point Ref(*x*). This reference point is chosen at random and all points in *S* have some probability of being selected as the reference point. Given that a point *xj* is closer to point *x*, the probability that this point is picked from several points *S* as a reference point for *x* is $\boldsymbol{P(\text{Ref}(x) = xj|S)}$ as measured by the distance function *dw*, where

$$d_w(x_{i,}x_j = \sum_{r=1}^{p} w_r^2 |x_{ir} - x_{jr}|, \tag{4.26}$$

w_r are the feature weights. Let's suppose that

$P(\text{Ref}(x) = xj|S) \propto k(dw(x, xj))$,

where *k* represent a kernel or a similarity function and its value becomes larger when *dw*(*x*,*xj*) is small. Assume it is

$$K(z) = \exp\left(-\frac{z}{\sigma}\right), \tag{4.27}$$

The reference point for *x* is selected from *S*, so the sum of *P*(Ref(*x*)= *xj*|*S*) for all *j* must be equal to 1. hence, it can be written as.

$$P\left(Ref(x) = x_j \middle| S^{-1}\right) = \frac{K(d_w(x_i, x_j))}{\sum_{j=1, j\neq 1}^{n} K(d_w(x_i, x_j))} \tag{4.28}$$

We applied the NCA on the generated features from the shoelace pattern feature extraction network, to discard the redundant features and select the most informative features from the generated features. The total features extracted from the feature creation phase are 2560 features. The application of the NCA reduces these features by the top 512 features. These 512 features are the final features that are fed to the machine learning classification algorithm. The steps below describe the implementation step of the neighborhood component analyzer.

Step 8: Utilize gf as input of NCA and generate 2560 weights.

Step 9: Create sorted indexes of these 2560 weights

Step 10: Select top 512 features using sorted indexes.

4.2.3. Classification

Classification in machine learning is a supervised learning approach in which the computer program learns from the available data to make new predictions. It is a procedure of categorizing a given set of data into classes from structured or unstructured data. The purpose of this classification step is to adopt a predictive modeling algorithm that will perform the task of approximating the mapping function from input variables to discrete output variables. The input variables to the adopted classifier are the selected features from the preceding phase. The main goal is to identify which class/category the new data will fall into. We adopted a support vector machine learning algorithm for mapping and categorizing the final feature vectors.

SVM is a supervised machine learning classification algorithm. An SVM model is simply a representation of different classes in a hyperplane in a multidimensional space. The hyperplanes are predicted iteratively by the SVM to determine the optimal hyperplane to minimize the error. This hyperplane is utilized on new datasets to classify them appropriately. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH). It is employed to classify a dataset in a 2-dimensional hyperplane as well as a multidimensional plane. Being a binary classifier, the training data set the hyperplane divides the training data set into two classes. To categorize multidimensional data, a multidimensional hyperplane with kernels must be employed to transform an input data space into the required form. it employs an approach called the kernel trick to transform a low dimensional input space and into a higher dimensional space. Simply put,

the kernel converts non-separable problems into separable problems by adding more dimensions to them. This makes the SVM a very powerful, flexible, and accurate algorithm. Some of the classes of kernels used by SVM are given below.

Linear Kernel: It can be used as a dot product between any two observations that have a linear relationship. The mathematical representation of the linear kernel is given below.

$$K(x, x_i) = sum(x * x_i) \tag{4.29}$$

The above formula that the product between two vectors says x & xi is the sum of the multiplication of each pair of input values.

Polynomial Kernel: Polynomial kernel is a more comprehensive form of the linear kernel. It distinguishes complex or non-linear input space. below is the mathematical formula of a polynomial kernel:

$$K(X, X_i) = 1 + sum(X + X_i)\text{^}d \tag{4.30}$$

Where d is the degree of the polynomial which we need to be manually specified in the learning algorithm.

Radial Basis Function (RBF) Kernel: Radial Basis Function kernel is commonly employed in SVM classification to map input space in indefinite dimensional space. The mathematical formula is expressed below.

$$K(x, x_i) = \exp(-\gamma * sum(x * x_i^2)) \tag{4.31}$$

Where γ, takes a value between 0 and 1. It also needs to be specified manually in the learning algorithm. A good default value for γ is 0.1.

As we implemented SVM can be implemented nonlinear data just like for linearly separable data by using kernels.

This classification phase is the last stage of the proposed ShoePatNet23. Cubic SVM was used to classify the selected 512 features, using a 10-fold cross-validation technique. The parameters of the employed cubic SVM classifier are given in below.

Kernel: 3rd-degree polynomial (Cubic),

Kernel scale: Automatic,

Box constraint level: 1,

Coding: One-vs-All,

Standardize: True,

Validation: 10-fold cross-validation.

The last step (Step 11) of the presented ShoePatNet23 model is given in this section.

Step 11: The results were obtained by classifying with Cubic SVM.

5. RESULTS AND DISCUSSION

In this section, we present the findings from the experiments. The findings are the result that we obtained using the proposed model (ShoePatNet23). The section also discusses the obtained results in detail. findings from the experiment.

5.1. Experimental Result

This section presents the results of the proposed ShoePatNet23 after it was tested with two speech datasets are. After proper preparation of the dataset, all the experiment was conducted with a simply configured desktop computer. The computer was used to program/implement the proposed ShoePatNet23. The computer has 32 GB memory, an intel i9 3.60 GHz processor, and a 512 GB solid-state disk. There is no adoption of any parallel programming technique in the implementation of our proposed ShoePatNet23 since it is a feedforward method. In this regard, ShoePatNet23 can be programmed on simply configured computers and does not require extensive or very powerful computing power. We also gathered two new speech datasets to test the proposed ShoePatNet23. It is worth noting that, ShoePatNet23 is a parametric learning model the parameters of which are also tabulated in Table 5.1 below.

Table 5.1. Parameters for ShoePatNet23

Step	Parameters
Speech segmentation	We divided speeches into non-overlapping widows/segments with a length of five seconds
TQWT	Two types of TQWT have been used in this work. The used parameters are; Q=[1,4], r=[3], J=[8,12]
Shoelace pattern	Signum, upper ternary, and lower ternary binary feature extraction functions have been used. Half of the standard deviation of the signal has been utilized as the threshold value of the ternary functions.
Chi2 selector	512 features are selected from the generated 3072 features
Loss value calculation	3rd-degree polynomial kernel SVM with five-fold cross-validation
Feature creation	The top five feature vectors are merged and 2560 features are created

Feature selection	The best 512 features from 2560 are selected by NCA. The parameters of the NCA are; verbosity level indicator is 0 (no convergence summary), the solver is stochastic gradient descend.
Classification	Cubic (3rd-degree polynomial kernel) SVM with 10-fold cross-validation. The parameters of the Cubic SVM are given in Section 3.3.

Applying the parameters in the above table(see Table 2), the model effectively classified the speech signal and yielded impressive results. Confusion matrices were used to further explain the result of the experiment is clear and clean terms as tabulated in Tables 5.2 and 5.3. Furthermore, other results (precision, recall, F1-score) have all been listed in these confusion matrices.

Table 5.2. Confusion matrix of the result of Turkish dataset

Actual class	Predicted class		
	Boring	Interesting	Neutral
Boring	2374	27	72
Interesting	12	2350	35
Neutral	55	54	2122
Recall (%)	96	98.04	95.11
Precision (%)	97.26	96.64	95.20
F1-score (%)	96.62	97.35	95.16

Table 5.3. Confusion matrix of the English dataset

Actual class	Predicted class		
	Boring	Interesting	Neutral
Boring	2760	13	92
Interesting	18	3275	153
Neutral	63	141	3026
Recall (%)	96.34	95.04	93.68
Precision (%)	97.15	95.51	92.51
F1-score (%)	96.74	95.27	93.09

Category-wise recall(accuracies), precision, and F1-score are given in Tables 5.2 and 5.3 above, in addition, we tabulated the overall results in Table 5.4. below where we compared the different performance metrics from the two datasets.

Table 5.4. The overall accuracy (%) of the ShoePatNet

Metric	Dataset in Turkish	Dataset in English
Accuracy	96.41	94.97
Precision	96.37	95.06
Recall	96.38	95.02
F1-score	96.38	95.04
Geometric mean	96.38	95.01
Cohen's kappa	94.61	92.43

5.2. Discussion

This section discusses the results that were obtained after testing the proposed model (ShoePatNet23) with speech emotion datasets. As stated in the sections above, the model has been tested on a Turkish speech dataset and an English speech dataset. Therefore we discuss the results obtained when the model was tested on the two datasets separately. The classification accuracy rates of each feature vector are denoted in Figure 5.1 below.

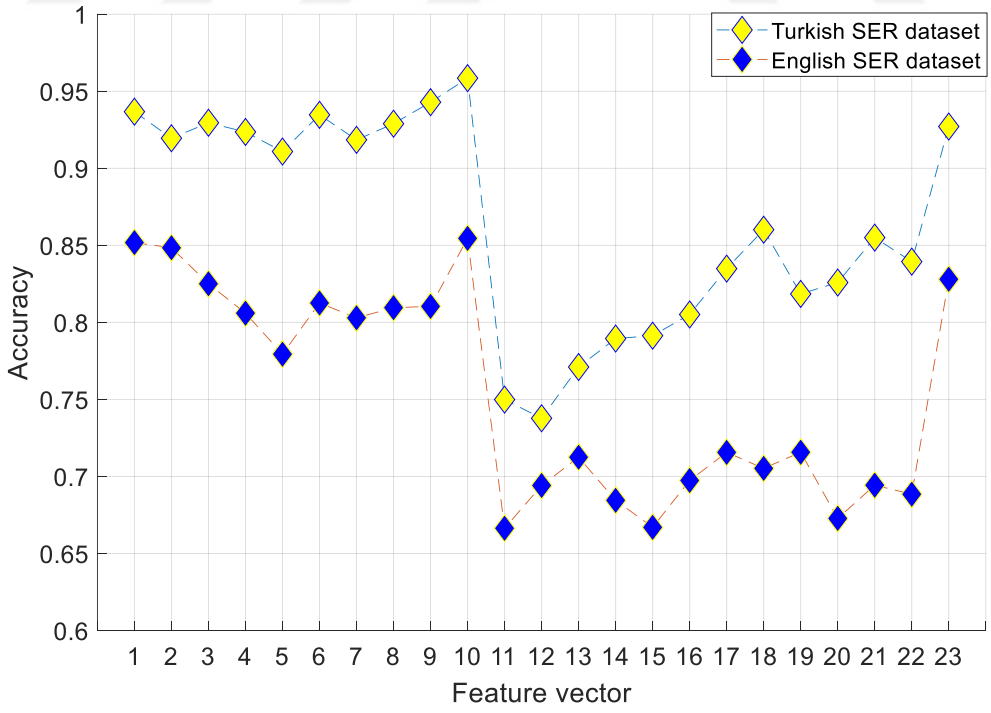


Figure 5.1. Accuracies (in %) of each feature vector based on the used dataset

As can be seen from Figure 5.1, the 10th feature vector gives the best result for both English and Turkish datasets, hence, becomes the best feature vector. 10th feature vector produced prediction accuracy of 95.85% on the Turkish SER dataset and 85.45% on the English SER dataset. For the Turkish SER dataset, the top five feature vectors are 10th, 9th, 1st, 6th, and 3rd feature vectors respectively, while 10th, 1st, 2nd, 23rd, and 3rd feature vectors are the top five performing feature vectors for the English SER dataset. But after concatenating the features and further applying a feature selection method of NCA, the best accuracies are increased from 95.85% to 96.41% for the Turkish SER dataset and from 85.45% to 94.97% for the English SER dataset. This demonstrates a drastic increase in the performance of the model on the English dataset. furthermore, the presented ShoePatNet23 produced an accuracy of over 65% for all feature vectors. Low oscillatory sub-bands also proved to be more effective (2nd – 10th feature vectors) than high oscillatory sub-bands (11th – 23rd feature vectors) for the two datasets that were used.

Cubic SVM classifier was used during the classification and computation of error values. Before selecting the Cubic SVM as the best appropriate classifier, the MATLAB classification learner toolbox was used to compare the performance of different classifiers. Some of the tested classifiers are decision tree (DT), linear discriminant (LD), quadratic discriminant (QD), naïve Bayes (NB), linear SVM (LSVM), Quadratic SVM (QSVM), Cubic SVM (CSVM), Gaussian SVM (GSVM), k nearest neighbor (kNN) and bagged tree (BT). The performance results for these classifiers are presented in Figure 5.2 below.

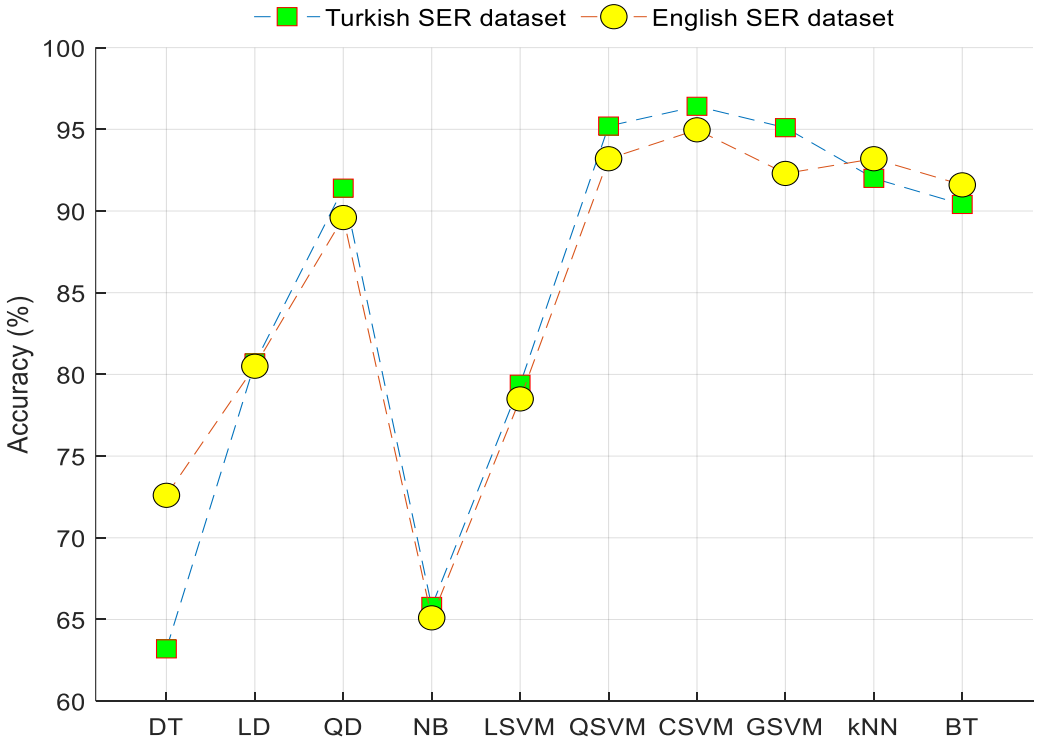


Figure 5.2. Calculated accuracies from ten classifiers

According to Figure 5.2 above, Cubic SVM (CSVM) is the best classifier. Therefore, we employed it in the proposed model (ShoePatNet23).

This research presented the benefits below, we also try to highlight the limitations of the research.

- The collection of two new speech emotion recognition datasets. This contributes to creating robust Speech emotion recognition databases in availability as well as in different languages (English and Turkish).
- The research also introduced a novel technique of local feature generation that is inspired by the patterns of how shoelaces are knotted and the mathematical concepts of graph theory. hence, the name shoelace pattern. It is supported further by implementing three binary feature extraction kernels, to create local histogram-based features.
- The entire network is a hand-modeled learning network. Our proposed ShoePatNet23 is a cognitive method and also tries to model deep learning architectures in its function. ShoePatNet23 proves to be an effective model for speech emotion recognition as demonstrated by its high classification accuracies of 94.97% and 96.41% on the English SER and Turkish SER datasets respectively.
- The research presents a contribution toward the development of new self-control applications to ensure the quality of learning in distance education systems. This research is a step toward achieving an interactive distance learning system and also has application in many fields.
- The model is a lightweight learning network that does not require very powerful computing resources.
- The overall success of the ShoePatNet23 is demonstrated with two datasets to demonstrate its versatility.
- To obtain spontaneous action, speeches are divided into non-overlapping blocks with a length of five seconds.

Limitations:

- Larger datasets can be collected.

The number of languages can be increased.

6. CONCLUSIONS

The distance learning system will continue to be adopted by many institutions of learning. It's a system that ensures the continuation of learning during the Covid-19 pandemic era. Many have realized its benefits beyond its role during the Covid period and will continue to adopt it post Covid as well. Therefore there is a need for ensuring the quality of learning through monitoring. To make distance education advantageous, smart systems should be developed with the data obtained from distance education platforms. This research focuses on speech processing and two datasets were collected for this purpose to detect the emotions of the instructors/lecturers during teaching. These datasets were collected from individuals that were asked to present a lecture manual and instructed to do so in three phases. The first is to present in a way and manner which would make the lecture interesting to students, then they are instructed to do the same for boring lecture and a neutral lecture (one that is neither boring nor interesting). In each case, they were to induce the necessary emotions. To demonstrate the general effectiveness of ShoePatNet23, the dataset was collected in two languages (Turkish and English). The proposed ShoePatNet23 generated 23 feature vectors from which it intelligently selects the top five feature vectors that will maximize the classification output.

The primary purpose of this research is to detect the emotional state of the instructors/lecturers while he/she is teaching from their voices. Our novel machine learning network is presented using the proposed shoelace pattern and it is named ShoePatNet23. The ShoePatNet23 is applied to the collected datasets producing 94.97% and 96.41% classification accuracies on the collected English and Turkish datasets respectively. The obtained results and findings demonstrate that ShoePatNet23 is a very successful success for the SER and can be applied to solve a real-world problem.

In near future, new self-control tools can be developed using the ShoePatNet23 and these tools can be installed to distance education platforms to evaluate lecturers' performance.
The contributions of our work to knowledge are listed below.

- This research contributes to progress in research on educational technologies.
- We introduced a novel hand-crafted feature extraction network which we called ShoePatNet23 which modeled a shoelace pattern. The shoelace pattern is a mathematical concept that we chose to adopt as a feature generation method. The proposed ShoePatNet23 calculates loss values in the feature extraction phase and uses these loss values to select the most valuable/appropriate signal for feature extraction.
- To solve the parameter tuning problem, high oscillatory and low oscillatory TQWT have been used to generate sub-bands before applying our ShoePatNet23 to select the most appropriate sub-band to generate features.

In this respect, the proposed ShoePatNet23 is a cognitive learning model.

RECOMMENDATIONS

Speech emotion recognition is an important pattern recognition task. Just like all machine learning and deep learning tasks, the success of the proposed model depends on a suitable dataset. This is why we collected a new dataset that is tailored towards the application of speech emotion recognition in classroom settings. To continue to improve on this work, we recommend that the amount of the data should be increased.

We also recommend that the services of professional actors may be employed in recording an induced speech dataset. However, we believe our dataset is more suitable for real-life modeling as the participants are given the free will to express themselves in ways and manner they feel are more suited to the category of lectures.

The best way to acquire this dataset is to put recording gadgets in an online learning environment to get a direct recording of a live class session. This makes it more natural, though may be very difficult to achieve but possible.

REFERENCES

[1] S. Dhawan, "Online learning: A panacea in the time of COVID-19 crisis," *Journal of Educational Technology Systems,* vol. 49, pp. 5-22, 2020.

[2] J. Nworie, "Beyond COVID-19: What's next for online teaching and learning in higher education," ed, 2020.

[3] Y. Liu, M. Zhang, X. Zhao, and F. Jia, "Fostering EFL/ESL Students' Language Achievement: The Role of Teachers' Enthusiasm and Classroom Enjoyment," *Frontiers in Psychology,* vol. 12, 2021.

[4] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems,* vol. 211, p. 106547, 2021.

[5] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control,* vol. 59, p. 101894, 2020.

[6] U. Asiya and V. Kiran, "Speech Emotion Recognition-A Deep Learning Approach," in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC)*, 2021, pp. 867-871.

[7] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 3020-3024.

[8] M. Laghari, M. J. Tahir, A. Azeem, W. Riaz, and Y. Zhou, "Robust Speech Emotion Recognition for Sindhi Language based on Deep Convolutional Neural Network," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2021, pp. 543-548.

[9] F. T. Villavicencio and A. B. Bernardo, "Positive academic emotions moderate the relationship between self-regulation and academic achievement," *British Journal of Educational Psychology,* vol. 83, pp. 329-340, 2013.

[10] A. C. Frenzel, R. Pekrun, and T. Goetz, "Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms," *Learning and Instruction,* vol. 17, pp. 478-493, 2007.

[11] M. Ainley, M. Corrigan, and N. Richardson, "Students, tasks, and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts," *Learning and Instruction,* vol. 15, pp. 433-447, 2005.

[12] T. Dalgleish and M. Power, *Handbook of cognition and emotion*: John Wiley & Sons, 2000.

[13] R. Pekrun, T. Goetz, W. Titz, and R. P. Perry, "Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research," *Educational Psychologist,* vol. 37, pp. 91-105, 2002.

[14] R. Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Educational psychology review,* vol. 18, pp. 315-341, 2006.

[15] R. Pekrun, A. J. Elliot, and M. A. Maier, "Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance," *Journal of Educational Psychology,* vol. 101, p. 115, 2009.

[16] R. Pekrun, "The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators," *Applied psychology,* vol. 41, pp. 359-376, 1992.

[17] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, and X. Xu, "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," *Neurocomputing*, vol. 448, pp. 238-248, 2021.

[18] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021.

[19] Q. Chen and G. Huang, "A novel dual attention-based BLSTM with hybrid features in speech emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104277, 2021.

[20] N. Londhe, M. Ahirwal, and P. Lodha, "Machine learning paradigms for speech recognition of an Indian dialect," in *2016 international conference on communication and signal processing (ICCSP)*, 2016, pp. 0780-0786.

[21] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social media and machine learning*, ed: IntechOpen, 2019.

[22] C. Cai, Y. Xu, D. Ke, and K. Su, "A fast learning method for multilayer perceptrons in automatic speech recognition systems," *Journal of Robotics*, vol. 2015, 2015.

[23] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22-35, 2019.

[24] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, pp. 101-108, 2012.

[25] H. K. Palo, M. N. Mohanty, and M. Chandra, "New features for emotional speech recognition," in *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*, 2015, pp. 424-429.

[26] M. M. Javidi and E. F. Roshan, "Speech emotion recognition by using combinations of C5. 0, neural network (NN), and support vector machines (SVM) classification methods," *Journal of mathematics and computer Science*, vol. 6, pp. 191-200, 2013.

[27] H. K. Palo, M. Chandra, and M. N. Mohanty, "Emotion recognition using MLP and GMM for Oriya language," *International Journal of Computational Vision and Robotics*, vol. 7, pp. 426-442, 2017.

[28] S. Motamed, S. Setayeshi, and A. Rabiee, "Speech emotion recognition based on a modified brain emotional learning model," *Biologically Inspired Cognitive Architectures*, vol. 19, pp. 32-38, 2017.

[29] B.-C. Chiou and C.-P. Chen, "Feature space dimension reduction in speech emotion recognition using support vector machine," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1-6.

[30] A. Jacob, "Modelling speech emotion recognition using logistic regression and decision trees," *International Journal of Speech Technology*, vol. 20, pp. 897-905, 2017.

[31] H. K. Palo and M. N. Mohanty, "Comparative analysis of neural networks for speech emotion recognition," *Int. J. Eng. Technol*, vol. 7, pp. 111-126, 2018.

[32] H. K. Palo, M. N. Mohanty, and M. Chandra, "Efficient feature combination techniques for emotional speech classification," *International Journal of speech technology*, vol. 19, pp. 135-150, 2016.

[33] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37-44, 2019.

[34] S. Prasomphan, "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram," in *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2015, pp. 73-76.

[35] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing*, vol. 22, pp. 1154-1160, 2012.

[36] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016, pp. 1080-1084.

[37] T. Rajisha, A. Sunija, and K. Riyas, "Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM," *Procedia Technology*, vol. 24, pp. 1097-1104, 2016.

[38] B. Yu, H. Li, and C. Fang, "Speech Emotion Recognition based on Optimized Support Vector Machine," *J. Softw.*, vol. 7, pp. 2726-2733, 2012.

[39] F. Chenchah and Z. Lachiri, "Speech emotion recognition in acted and spontaneous context," *Procedia Computer Science*, vol. 39, pp. 139-145, 2014.

[40] M. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech-based human emotion recognition using MFCC," in *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2017, pp. 2257-2260.

[41] N. Kamaruddin and A. Wahab, "Driver behavior analysis through speech emotion understanding," in *2010 IEEE Intelligent vehicles symposium*, 2010, pp. 238-243.

[42] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of Teager energy operator and MFCC," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2017, pp. 1-5.

[43] T. Iliou and C.-N. Anagnostopoulos, "SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study," in *2010 Fifth International Conference on Digital Telecommunications*, 2010, pp. 1-6.

[44] G. N. Peerzada, R. Deshmukh, and S. Waghmare, "A review: Speech emotion recognition," *Int. J. Comput. Sci. Eng*, vol. 6, pp. 400-402, 2018.

[45] K. Khanchandani and M. A. Hussain, "Emotion recognition using multilayer perceptron and generalized feed-forward neural network," 2009.

[46] A. P. Reddy and V. Vijayarajan, "Audio compression with multi-algorithm fusion and its impact in speech emotion recognition," *International Journal of Speech Technology*, vol. 23, pp. 277-285, 2020.

[47] M. Imani and G. A. Montazer, "A survey of emotion recognition methods with emphasis on E-Learning environments," *Journal of Network and Computer Applications*, vol. 147, p. 102423, 2019.

[48] K. Chen, G. Yue, F. Yu, Y. Shen, and A. Zhu, "Research on speech emotion recognition system in e-learning," in *International Conference on Computational Science*, 2007, pp. 555-558.

[49] A. Ray and A. Chakrabarti, "Design and implementation of affective e-learning strategy based on facial emotion recognition," in *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*, 2012, pp. 613-622.

[50] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, technology, design, and learning*, ed: Elsevier, 2016, pp. 27-46.

[51] J. Kreiman and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*: John Wiley & Sons, 2011.

[52] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, p. 614, 1996.

[53] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion,* vol. 15, pp. 123-148, 1991.

[54] R. G. Kamiloğlu, A. H. Fischer, and D. A. Sauter, "Good vibrations: A review of vocal expressions of positive emotions," *Psychonomic bulletin & review,* vol. 27, pp. 237-265, 2020.

[55] A. H. Gardiner, "The theory of speech and language," 1932.

[56] M. Cabanac, "What is emotion?," *Behavioural processes,* vol. 60, pp. 69-83, 2002.

[57] N. K. Denzin, *On understanding emotion*: Routledge, 2017.

[58] P. G. Zimmermann, "Beyond usability: measuring aspects of user experience," ETH Zurich, 2008.

[59] R. O. Stanley and G. D. Burrows, "Varieties and functions of human emotion," *Emotions at work: Theory, research, and applications in management,* pp. 3-19, 2001.

[60] D. Keltner and J. S. Lerner, "Emotion," 2010.

[61] E. A. Phelps, "Emotion and cognition: insights from studies of the human amygdala," *Annu. Rev. Psychol.,* vol. 57, pp. 27-53, 2006.

[62] R. D. Ray and D. H. Zald, "Anatomical insights into the interaction of emotion and cognition in the prefrontal cortex," *Neuroscience & Biobehavioral Reviews,* vol. 36, pp. 479-501, 2012.

[63] G. Pitcher, "Emotion," *Mind,* pp. 326-346, 1965.

[64] A. Öhman, A. Flykt, and F. Esteves, "Emotion drive attention: detecting the snake in the grass," *Journal of experimental psychology: general,* vol. 130, p. 466, 2001.

[65] S. B. Most, M. M. Chun, D. M. Widders, and D. H. Zald, "Attentional rubbernecking: Cognitive control and personality in emotion-induced blindness," *Psychonomic bulletin & review,* vol. 12, pp. 654-661, 2005.

[66] G. Matthews and A. Wells, "The cognitive science of attention and emotion," 1999.

[67] B. Knutson, G. E. Wimmer, C. M. Kuhnen, and P. Winkielman, "Nucleus accumbens activation mediates the influence of reward cues on financial risk-taking," *NeuroReport,* vol. 19, pp. 509-513, 2008.

[68] J. Panksepp, "Brain emotional circuits and psychopathologies," in *Emotions and psychopathology*, ed: Springer, 1988, pp. 37-76.

[69] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion," *American psychologist,* vol. 46, p. 819, 1991.

[70] A. á Ortony, G. á Clore, and A. á Collins, "The cognitive structure of emotions," ed: Cambridge University Press, Cambridge, UK, 1988.

[71] P. Ekman, "8c Friesen, WV (1978). The Facial Action Coding System: A technique for the measurement of facial movement. Palo Alto," ed: CA: Consulting Psychologists Press.

[72] C. Darwin, "The expression of the emotions in man and animals (1872)," *The Portable Darwin,* pp. 364-393, 1993.

[73] K. O'regan, "Emotion and e-learning," *Journal of Asynchronous learning networks,* vol. 7, pp. 78-92, 2003.

[74] A. R. Damasio, "Descartes' error: Emotion, rationality and the human brain," *New York: Putnam,* vol. 352, 1994.

[75] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology,* vol. 39, p. 1161, 1980.

[76] T. Hascher, "Learning and emotion: Perspectives for theory and research," *European Educational Research Journal,* vol. 9, pp. 13-28, 2010.

[77] P. A. Schutz, L. P. Aultman, and M. R. Williams-Johnson, "Educational psychology perspectives on teachers' emotions," in *Advances in teacher emotion research*, ed: Springer, 2009, pp. 195-212.

[78] G. H. Bower, "Mood and memory," *American psychologist*, vol. 36, p. 129, 1981.

[79] N. Schwarz, *Feelings as information: Informational and motivational functions of affective states*: The Guilford Press, 1990.

[80] B. Kort, R. Reilly, and R. W. Picard, "An affective model of the interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in *Proceedings IEEE international conference on advanced learning technologies*, 2001, pp. 43-46.

[81] R. Calvo, D. Peters, and L. Peters, "Two approaches for the design of affective computing environments for education," in *AIED 2009: 14th International Conference on Artificial Intelligence in Education Workshops Proceedings*, 2009.

[82] K. R. Scherer, "What are emotions? And how can they be measured? " *Social science information*, vol. 44, pp. 695-729, 2005.

[83] M. Feidakis, T. Daradoumis, and S. Caballé, "Emotion measurement in intelligent tutoring systems: what, when and how to measure," in *2011 Third International Conference on Intelligent Networking and Collaborative Systems*, 2011, pp. 807-812.

[84] M. Wong, "Emotion assessment in the evaluation of affective interfaces," *Neuron*, vol. 65, p. 293, 2006.

[85] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *arXiv preprint cs/0212032*, 2002.

[86] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273-297, 1995.

[87] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for information science*, vol. 41, pp. 391-407, 1990.

[88] G. Leshed and J. J. Kaye, "Understanding how bloggers feel: recognizing affect in blog posts," in *CHI'06 extended abstracts on Human factors in computing systems*, 2006, pp. 1019-1024.

[89] S. Caballé, À. Lapedriza, D. Masip, F. Xhafa, and A. Abraham, "Enabling Automatic Just-in-time Evaluation of In-class Discussions in On-line Collaborative Learning Practices," *J. Digit. Inf. Manag.*, vol. 8, pp. 323-330, 2010.

[90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "WordNet: An online lexical database. 1990," *Int. J. Lexicography*, vol. 3.

[91] X. Carreras, I. Chao, L. Padró, and M. Padró, "FreeLing: An Open-Source Suite of Language Analyzers," in *LREC*, 2004, pp. 239-242.

[92] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, pp. 1162-1181, 2006.

[93] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *INTERSPEECH*, 2002.

[94] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk, "Promise-a procedure for multimodal interactive system evaluation," in *Multimodal Resources and Multimodal Systems Evaluation Workshop Program Saturday, June 1, 2002*, 2002, p. 14.

[95] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in an intelligent household robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, 2010, pp. 537-541.

[96] A. Das, P. Acharjee, L. K. Thakuria, and P. Talukdar, "A brief study on speech emotion recognition," *International Journal of Scientific &Engineering Research (IJSER),* vol. 5, pp. 339-343, 2014.

[97] D. A. Sauter, C. Panattoni, and F. Happé, "Children's recognition of emotions from vocal cues," *British Journal of Developmental Psychology,* vol. 31, pp. 97-113, 2013.

[98] P. N. Juslin and K. R. Scherer, *Vocal expression of affect*: Oxford University Press, 2005.

[99] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication,* vol. 40, pp. 227-256, 2003.

[100] K. R. Scherer, "Vocal affect expression: a review and a model for future research," *Psychological Bulletin,* vol. 99, p. 143, 1986.

[101] K. R. Scherer, T. Johnstone, and G. Klasmeyer, *Vocal expression of emotion*: Oxford University Press, 2003.

[102] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological Bulletin,* vol. 129, p. 770, 2003.

[103] B. Polster, "What is the best way to lace your shoes?," *Nature,* vol. 420, pp. 476-476, 2002.

[104] I. W. Selesnick, "Wavelet transform with tunable Q-factor," *IEEE transactions on signal processing,* vol. 59, pp. 3560-3575, 2011.

[105] I. Bayram and I. W. Selesnick, "Frequency-domain design of overcomplete rational-dilation wavelet transforms," *IEEE transactions on signal processing,* vol. 57, pp. 2957-2972, 2009.