Department of Physical Science

Faculty of Applied Science

Declaration

I hereby declare that the research project submitted for evaluation of course module IT4216 leading to the award of a Bachelor of Science Honors in Information Technology is entirely my work, and the contents taken from the work of others have been cited and acknowledged within the text. This proposal has not been submitted for any degree at this University or any other institution.

……..
L.P.S.Anjana
2018/ICT/01

I recommend the project to be carried out by the student, 2018/ICT/01

11/03/2024

…….. ……..

Dr. S. Kirushanth
Department of Physical Science,
Senior Lecturer,
Faculty of Applied Science.

Date

Proposal for the Final Year Research Project

B.Sc. (Honors) in Information Technology

Title: Emotion Detection from Voice using Deep Learning Algorithms

Name: L.P.S.Anjana

Registration Number: 2018/ICT/01

Supervisor(s) Name: Dr. S. Kirushanth (Senior Lecturer, Department of Physical Science)

Keywords: Emotion, Speech analysis, Voice identification, criminal law, emotion recognition, feature extraction, speech recognition

1. **Background**

Emotion serves as a fundamental pillar in our daily human interactions, contributing significantly to both our rational and intelligent decision-making processes. It equips us with the ability to empathize and understand the emotions of others, by allowing us to express our feelings and respond to theirs. Studies have highlighted the substantial impact of emotions on human social dynamics, demonstrating how the display of emotions can reveal a wealth of information about an individual's mental state. This understanding has given birth to a new field of research known as automatic emotion recognition, which is dedicated to deciphering and retrieving the specific emotions expressed. Previous research has investigated various methods for identifying emotional states, including the analysis of facial expressions [1], speech [2], and physiological signals [3] among others.

Facial expressions play a huge part in emotions. We cannot always capture the facial expressions. facial expressions are not reflected in situations like phone calls or voice recordings. In most criminal activities there is a call to someone. Analyzing those calls and getting the emotions of the calling person is helpful in detecting that crime (forensics) [4].

Speech signals offer several inherent benefits that make them an ideal resource for affective computing. For example, speech signals can usually be acquired more readily and economically compared to many other biological signals (e.g., electrocardiogram). This has led to a predominant interest among researchers in the field of speech emotion recognition (SER). SER focuses on identifying the hidden emotional state of a speaker through their vocal cues. Over recent years, this area has seen a growing surge in research attention.

Detecting human emotions has many applications, such as in robotic interfaces, audio surveillance, web-based E-learning, commercial endeavors, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, and more. Particularly in the context of classroom orchestration or E-learning, insight into a student's emotional state can significantly contribute to improving the quality of instruction. For instance, a teacher could utilize SER to determine appropriate subjects to teach and devise effective strategies to manage emotional dynamics within the learning environment. This highlights the importance of considering the emotional state of learners in a classroom setting.

2. Literature Review

Variations in the autonomic nervous system can influence a person's speech, and affective technologies can interpret this data to identify emotions. For instance, emotions such as fear, anger, or joy often result in speech that is louder, faster and encompasses a higher and broader pitch range. On the other hand, emotions like sadness or fatigue typically lead to speech that is slower and lower in pitch [5]. Certain emotions, such as anger or approval, have been observed to be more readily identifiable through computational methods [6].

Emotional speech processing technologies recognize the user's emotional state using computational analysis of speech features. Through pattern recognition techniques, vocal parameters and prosodic characteristics like pitch variations and speech rate can be scrutinized [6], [7].

Table 1 presents an overview of the key parameters to be examined in digital speech or voice recordings during the process of feature extraction.

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

Table 1. Emotions and Speech Parameters (from Murray and Arnott, 1993)[8]

Speech analysis is an effective method of identifying affective state. Some research reports an average accuracy of 70% to 80% [9], [10] in emotion detection through speech analysis, which surpasses the average human accuracy of approximately 60% [6]. However, it falls short of the accuracy achieved by other emotion detection systems that measure physiological states or facial expressions [11].

Figure 1 illustrates the two components of emotion recognition based on speech: the concurrent analysis of speech content and speech features (see Table 1). The semantic aspect of this analysis involves counting the occurrences of words with emotional connotations. A fundamental categorization includes 'positive' versus 'negative' mental states.

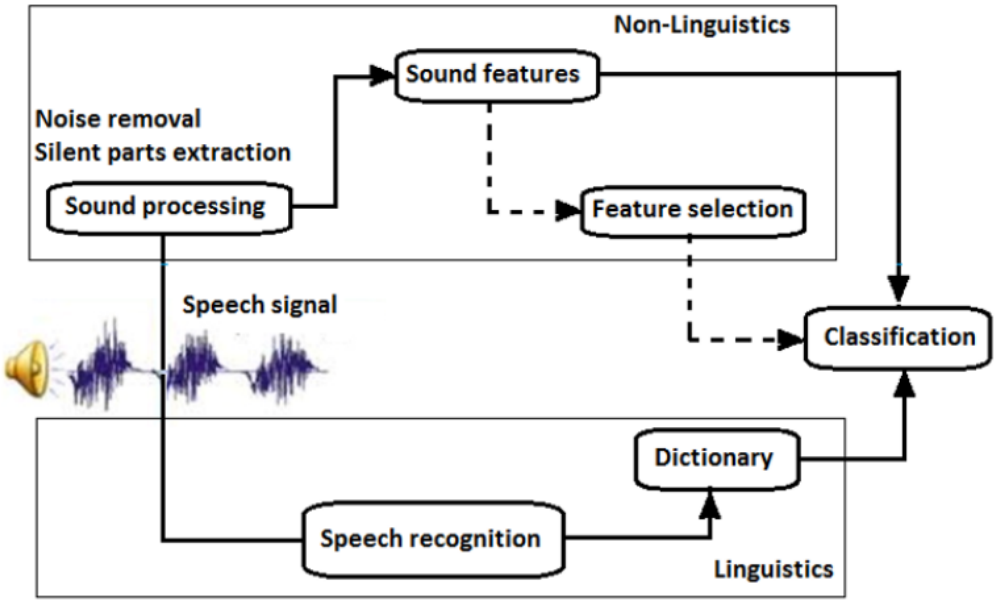

Figure 1. Speech-based emotion detection (Anagnostopouloset al., 2012) [11]

Emotion recognition differs from, and indeed complements, speech recognition. In contrast to speech recognition, where researchers develop algorithms and applications that automatically generate thousands of hours of transcribed speech, there isn't a standardized or unified approach for emotion detection and analysis from the human voice [12]. However, there is a general agreement on the top six most significant emotions to be recognized, often referred to as 'the big six' (see Figure 2). This approach has been greatly propelled by the extensive analysis conducted by Google Research in the Audio Set project [13]. The analysis of over two million videos from YouTube channels yielded a vast set of over 600 audio classes (audio events). The entire process of analysis hinges on feature extraction, detection, and recognition, using Mel-frequency cepstral coefficients (MFCC) based acoustic features and a General Mixture Model (GMM) based classifier.

Figure 2. Plutchik's wheel of emotions simplified. [14]

The application of deep learning methods, specifically deep feed-forward neural networks such as Convolutional Neural Networks (CNN) and recurrent neural networks, represents a relatively novel approach [15]. While the outcomes of 'classic' methods for some of the 'big six' emotions were encouraging [11], the latest advancements involving deep CNN have been truly remarkable [16], [17], [18].

3. Aim and Objectives

The overall aim of this research is to develop and evaluate deep learning algorithms for emotion detection from voice data, with a focus on enhancing crime detection and forensic analysis.

To achieve this aim, outlined several objectives will be following.

Understanding the Role of Emotions in Criminal Behavior:

- Conduct a comprehensive literature review to explore the influence of emotions on human interactions, particularly in the context of criminal behavior.
- Investigate existing empirical research to establish a theoretical framework for understanding the role of emotions in crime scene analysis.

Exploring Methods for Emotion Recognition:

- Review various methods for emotion recognition, including facial expressions, physiological responses, and speech analysis.
- Examine the effectiveness of speech analysis in capturing emotional states, considering its advantages over other modalities.

Evaluation of Speech Signals for Emotion Recognition:

- Evaluate the suitability of speech signals as a source for emotion recognition, emphasizing their rich emotional cues and accessibility.
- Conduct experiments to demonstrate the effectiveness of speech signals in accurately identifying emotional states.

Highlighting the Importance of SER in Crime Detection:

- Investigate the practical applications of Speech Emotion Recognition (SER) in crime detection, forensic analysis, and security systems.
- Explore the potential impact of SER on improving crime detection rates and enhancing public safety.

Assessment of SER Effectiveness in Crime Detection:

- Rigorously analyze the effectiveness of speech analysis in identifying emotional states, particularly in the context of crime detection.
- Utilize empirical studies and data analysis to assess the accuracy and reliability of SER in detecting emotional cues related to criminal behavior.

Exploring Deep Learning Techniques for Emotion Detection:

- Investigate the potential of deep learning methods, such as Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs), in improving emotion recognition from speech.
- Develop and test deep learning models tailored for emotion detection, aiming to enhance accuracy and efficiency in crime scene analysis.

4. Methodology

The methodology section outlines the systematic approach that will be undertaken to achieve the research objectives of emotion detection from voice data. This section details the steps involved in data collection, preprocessing, model development, evaluation, and analysis.

4.1 Data Collection:

The dataset utilized for this research was sourced from two primary repositories: Mozilla Common Voice [19] for gender and age data, and the RAVDESS database [20] for emotion data. These datasets comprised audio recordings from diverse sources, providing a rich source of information for training and testing models. The dataset consisted of 20 statistical features extracted through Frequency Spectrum Analysis using the R programming language, accompanied by labels denoting gender, age, and emotion.

4.2 Data Preprocessing:

Data preprocessing is vital for ensuring the quality and consistency of the dataset. Preprocessing tasks include handling missing values, removing outliers, and standardizing feature values. Additionally, the dataset will be split into training, validation, and testing sets to facilitate model development and evaluation. Feature normalization will be applied to ensure uniformity across features and enhance model performance.

4.3 Model Development:

Deep learning algorithms, specifically Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs), will be employed for emotion detection from voice data. The architecture of the models will be designed to accommodate the input features (20 statistical features) and predict the corresponding emotion. Training of the models will be conducted using the training dataset, with hyperparameter optimization aimed at improving model performance.

4.4 Evaluation:

The performance of the trained models will be assessed using the validation dataset to gauge their generalization ability and effectiveness in predicting gender, age, and emotion from voice data. Evaluation metrics such as accuracy, precision, recall, and F1-score will be employed to measure model performance. Comparative analysis will be conducted against baseline methods and state-of-the-art approaches to benchmark the proposed models.

4.5 Analysis:

The results of the model evaluation will be analyzed to identify the strengths and weaknesses of the proposed approach. Interpretation of findings will be contextualized within the research objectives, with a focus on implications for crime detection and forensic analysis. Insights gained from the analysis will inform potential areas for improvement and future research directions in the field of emotion detection from voice using deep learning algorithms.

By following this methodology, the research aims to advance the state-of-the-art in emotion detection from voice data, with potential applications in crime detection and forensic analysis.

5. Work Plan

Activity	1st month			2nd month			3rd month			4th month			5th month			6th month			7th month		
Research Proposal																					
Literature Review/Data Cleaning																					
Design algorithms and models																					
Evaluation																					
Result/report																					

REFERENCES

[1] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom, "Facial emotion recognition using empirical mode decomposition," *Expert Syst Appl*, vol. 42, no. 3, pp. 1261–1277, Feb. 2015, doi: 10.1016/j.eswa.2014.08.049.

[2] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, Jan. 2018, doi: 10.1016/j.neucom.2017.07.050.

[3] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors," 2018, pp. 15–22. doi: 10.1007/978-3-319-60639-2_2.

[4] A. Gully, P. Harrison, V. Hughes, R. Rhodes, and J. Wormald, "How Voice Analysis Can Help Solve Crimes," *Front Young Minds*, vol. 10, Feb. 2022, doi: 10.3389/frym.2022.702664.

[5] C. Breazeal and L. Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech," *Auton Robots*, vol. 12, no. 1, pp. 83–104, 2002, doi: 10.1023/A:1013215010749.

[6] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, IEEE, pp. 1970–1973. doi: 10.1109/ICSLP.1996.608022.

[7] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, IEEE, pp. 240–243. doi: 10.1109/ASRU.2001.1034632.

[8] I. R. Murray and J. L. Arnott, “Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech,” *Comput Speech Lang*, vol. 22, no. 2, pp. 107–129, Apr. 2008, doi: 10.1016/j.csl.2007.06.001.

[9] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion Recognition in Spontaneous Speech Using GMMs.”

[10] S. Yacoub, S. Simske, X. Lin, and J. Burns, “Recognition of Emotions in Interactive Voice Response Systems,” 2003. [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/

[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artif Intell Rev*, vol. 43, no. 2, pp. 155–177, Feb. 2015, doi: 10.1007/s10462-012-9368-5.

[12] F. Weninger, M. Wöllmer, and B. Schuller, “Emotion Recognition in Naturalistic Speech and Language-A Survey,” in *Emotion Recognition*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, pp. 237–267. doi: 10.1002/9781118910566.ch10.

[13] “Alarge-scaledatasetofmanuallyannotatedaudioevents.” Accessed: Aug. 07, 2023. [Online]. Available: https://research.google.com/audioset/index.html

[14] R. Plutchik, “The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *Am Sci*, vol. 89, no. 4, pp. 344–350, 2001, [Online]. Available: http://www.jstor.org/stable/27857503

[15] A. Konar and A. Chakraborty, Eds., *Emotion Recognition*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015. doi: 10.1002/9781118910566.

[16] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and Recurrent Neural Networks,” in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, Dec. 2016, pp. 1–4. doi: 10.1109/APSIPA.2016.7820699.

[17] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, “A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks,” Jul. 2017.

[18] M. Neumann and N. T. Vu, “Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech,” Jun. 2017.

[19] “Mozilla Common Voice.” Accessed: May 29, 2021. [Online]. Available: https://smartlaboratory.org/ravdess/

[20] “RAVDESS database.” Accessed: May 29, 2021. [Online]. Available: https://commonvoice.mozilla.org/en