A black and purple background

Description automatically generated

A purple and blue waves on a black background

Description automatically generated

Word Count: XXXX

Date: DD/MM/YYYY

KV6003 – Individual Computing Project

Panagiotis Theodorou – W20020633

Supervisor: Dr Hua Mao

2nd Marker: Dr Nic Whitton

A Comparative Analysis of Machine Learning and Deep Learning Models

Speech Emotion Recognition (S.E.R)

**2023 - 2024**

**1.0 - Introduction**

**1.1 - Background**

**1.2 - Objectives**

**1.3 - Scope and Significance**

**2.0 - Literature Review**

**2.1 - Overview of Speech Emotion Recognition**

**2.2 - Studies of Models in Speech Emotion Recognition**

**2.2.1 - Deep Learning**

**2.2.2 - Machine Learning**

**2.2.3 - Ensemble Learning**

**2.3 - Challenges and Limitations**

**2.3.1 - Datasets**

**2.3.1.1 - CREMA-D**

**2.3.1.2 - RAVDESS**

**2.3.1.3 - SAVEE**

**2.3.1.4 - EMO-DB**

**2.3.2 - Emotions**

**2.3.3 - Authenticity**

**2.3.4 - Cultural and Gender biases**

**2.3.5 - Subjectivity and Variability**

**2.3.6 - Overfitting**

**3.0 - Methodology**

**3.1 - Feature Extraction**

**3.2 - Augmentation**

**3.3 - Performance**

**3.3.1 - Deep Learning Models**

**3.3.1.1 - ANN**

**3.3.1.2 - CNN**

**3.3.1.3 - MLP**

**3.3.1.4 - ResNet**

**3.3.1.5 - RNN+LSTM**

**3.3.1.6 - Transformer**

**3.3.2 - Machine Learning Models**

**3.3.2.1 - GMM**

**3.3.2.2 - HMM**

**3.3.2.3 - KNN**

**3.3.2.4 - RF**

**3.3.2.5 - SVM**

**3.3.3 - Ensemble Models**

**3.3.3.1 - Lorem Ipsum 1**

**3.3.3.2 - Lorem Ipsum 2**

**3.3.3.3 - Lorem Ipsum 3**

**3.3.3.4 - Lorem Ipsum 4**

**3.3.3.5 - Lorem Ipsum 5**

**3.4 - Lorem Ipsum**

**3.5 - Model Selection**

**4.0 - Own Approach**

**4.1 - Lorem Ipsum 1**

**4.2 - Lorem Ipsum 2**

**4.3 - Lorem Ipsum 3**

**4.4 - Lorem Ipsum 4**

**4.5 - Lorem Ipsum 5**

**5.0 - Implementation**

**5.1 - Lorem Ipsum 1**

**5.2 - Lorem Ipsum 2**

**5.3 - Lorem Ipsum 3**

**5.4 - Lorem Ipsum 4**

**5.5 - Lorem Ipsum 5**

**6.0 - Results**

**6.1 - Lorem Ipsum 1**

**6.2 - Lorem Ipsum 2**

**6.3 - Lorem Ipsum 3**

**6.4 - Lorem Ipsum 4**

**6.5 - Lorem Ipsum 5**

**7.0 - Discussion**

**7.1 - Lorem Ipsum 1**

**7.2 - Lorem Ipsum 2**

**7.3 - Lorem Ipsum 3**

**7.4 - Lorem Ipsum 4**

**7.5 - Lorem Ipsum 5**

**8.0 - Conclusion**

**8.1 - Lorem Ipsum 1**

**8.2 - Lorem Ipsum 2**

**8.3 - Lorem Ipsum 3**

**8.4 - Lorem Ipsum 4**

**8.5 - Lorem Ipsum 5**

**Title: Speech Emotion Recognition Using Machine Learning and Deep Learning Models**

**Abstract:**

Provide a concise summary of your report, including the purpose, methodology, key findings, and implications of your research.

# Introduction

## Background

## Objectives

## Scope and Significance

# Literature Review

## Overview of Speech Emotion Recognition

The Speech Emotion Recognition (SER) domain is a multidisciplinary field encompassing various key terms, concepts, and techniques to comprehend emotions conveyed through human speech (Wani et al., 2021). Paralinguistic information, such as gender, personality, emotion, aim, and state of mind, is integral in decoding the hidden significance of utterances within SER (Wani et al., 2021). Initial emphasis in classical automatic speech recognition systems, as highlighted in Wani et al.'s review, was on linguistic content, with less attention to paralinguistic information (Wani et al., 2021). However, the evolution of SER acknowledges the profound impact of meagreness in paralinguistic awareness on communication, especially in children, leading to substandard social skills and potential psychopathological manifestations (Wani et al., 2021). Modern SER systems aim to create coherent and human-like communication machines by comprehending paralinguistic data, especially emotional nuances, to enhance communication effectiveness.

Methodologies employed in SER encompass diverse approaches, including traditional classifiers like Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM), k-Nearest Neighbour (kNN), and contemporary techniques like Deep Neural Networks (DNN) (Wani et al., 2021). These illustrate the evolution of classification methods for emotion recognition. Feature extraction techniques in SER, as discussed in both articles, range from prosodic and spectral features to voice quality features and Teager Energy Operator (TEO)-based features, showcasing the comprehensive nature of information considered (Wani et al., 2021).

Jaiswal et al. (2020) outlined the processing steps involved in speech signal analysis, including framing, windowing, voice activity detection, normalization, and noise reduction. The significance of feature selection and dimension reduction techniques is underscored to mitigate challenges associated with high-dimensional data (Jaiswal et al., 2020). Global features, representing supra-segmental or long-term characteristics, and local features, capturing segmental or short-term dynamics, are critical components in feature extraction (Jaiswal et al., 2020). Various spectral features, such as Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), and Gammatone Frequency Cepstral Coefficients (GFCC), offer insights into the frequency domain characteristics of speech signals (Jaiswal et al., 2020). The importance of classifiers, both traditional and deep learning-based, is evident in the choice of models like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), k-Nearest Neighbour (kNN), and Deep Neural Networks (DNN) for accurate emotion classification (Jaiswal et al., 2020). These fundamental elements collectively contribute to the rich landscape of Speech Emotion Recognition, reflecting the interdisciplinary nature of the field, from linguistics and psychology to signal processing and machine learning.

## Studies of Models in Speech Emotion Recognition

### Deep Learning Models

Deep Learning models, as explored in the article by Fayek, Lech, and Cavedon (2017), demonstrate significant capabilities in Speech Emotion Recognition (SER). These models can effectively handle both static and dynamic classification problems, making them versatile for various applications. The employment of minimal speech processing and end-to-end deep learning architectures allows for the modelling of intra-utterance dynamics, which is crucial for recognising emotional states from speech. The benefits of using Deep Learning for SER include achieving state-of-the-art results on the IEMOCAP database for speaker-independent SER, demonstrating the models' ability to generalise across different speakers. Furthermore, these models present a simple pipeline and low latency in emotion recognition tasks, which are advantageous for real-time applications. The exploration of feed-forward and recurrent neural network architectures, along with their variants, illuminates their strengths and limitations in handling paralinguistic elements of speech, offering insights into the future development of more sophisticated and efficient SER systems (Fayek, H.M., Lech, M. & Cavedon, L., 2017).

Pandey et al. (2019) explored the use of deep learning techniques in recognizing emotions from speech, highlighting the significance of enabling machines not only to understand the content but also to interpret the emotional nuances of human speech. Unlike traditional machine learning models, which are limited in their ability to capture the complexities of emotional states, deep learning offers a more effective approach by automatically learning high-level features from data.

In their evaluation, Pandey et al. (2019) used "Weighted Accuracy" to compare different deep learning architectures and input features across four basic emotions: Neutral, Angry, Happy, and Sad. Their findings indicated that the best results were obtained using CNN, LSTM, and a hybrid model combining CNN and BLSTM. To ensure the robustness of their findings and mitigate any potential bias from a single dataset, they utilized a five-fold cross-validation approach with two datasets, Emo-DB and IEMOCAP. The hybrid model (CNN+BLSTM) with Mel-Frequency Cepstral Coefficients (MFCC) as the input feature achieved the highest accuracy (82.35%) on the Emo-DB dataset. This model, however, faced challenges in accurately classifying happy emotions, attributed to class imbalances and the arousal similarities between happy and angry utterances. Conversely, for the IEMOCAP dataset, the same hybrid model with a Mel-Spectrogram input feature performed best. Interestingly, it accurately classified "happy" emotions more effectively than others, which Pandey et al. (2019) suggest may be due to the more natural elicitation of emotions in the IEMOCAP dataset, despite an overall accuracy below 50%.

|  |
| --- |
|  |
| ***Fig:*** *Confusion Matrices of Pandey et al. (2019) evaluations of both datasets across 4 emotions in a five-fold validation.* |

### Machine Learning Models

### Ensemble Models

## Challenges and Limitations

### Datasets

When developing a Speech Emotion Recognition model in Artificial Intelligence, one faces several challenges related to the datasets used. The quality of the dataset is paramount; poor audio quality, background noise, or poorly labelled data can significantly hinder model performance. A model can only be as good as the data it learns from, making high-quality datasets essential for accurate emotion classification.

The size of the dataset is another critical factor. Training robust models requires large amounts of data to capture the variability in speech patterns. However, collecting and curating such vast datasets is resource-intensive and often impractical. Small datasets may lead to overfitting, where the model performs well on training data but poorly on unseen data.

Dataset metadata, including information about the speakers (such as age, gender, and native language), recording conditions, and emotional states, is crucial for understanding and contextualizing the data. Lack of detailed metadata can prevent the model from learning nuanced differences in speech that may be influenced by these factors.

Feature extraction is a challenge, as determining which features of the speech are most relevant for emotion recognition is not straightforward. The extraction and selection of features such as pitch, tone, speed, and pauses directly impact the model's ability to learn and classify emotions accurately.

The generalization complexity of datasets refers to their ability to represent real-world variability. Datasets often lack diversity, focusing on specific demographics or languages, which limits the model's ability to generalize across different populations and situations.

Lastly, datasets that do not include facial expressions or other non-verbal cues present a unique challenge. In real-world applications, emotional recognition often relies on a combination of verbal and non-verbal cues. Relying solely on speech data may limit the model's effectiveness in accurately classifying emotions, as it misses out on the rich context provided by facial expressions and body language.

Addressing these challenges requires careful dataset selection, augmentation strategies, and feature engineering to develop a Speech Emotion Recognition model capable of accurately classifying emotions across diverse conditions and populations.

#### CREMA-D

The CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) dataset, created by Cao et al. (2014), is a rich resource developed to enable detailed study and analysis in the field of speech emotion recognition. Comprising 7,442 video clips from 91 actors, including 48 men and 43 women, ranging in age from 20 to 74 years and representing a wide array of ethnic backgrounds, it is designed to reflect the diversity of emotional expression and perception. These actors were recorded expressing six basic emotions: happiness, sadness, anger, fear, disgust, and neutral, with variations in emotional intensity, providing a nuanced spectrum of emotional states for analysis.

Each clip in the dataset features a spoken dialogue designed to fit within a neutral context, allowing the emotional expression to be the focal point, free from biases that might arise from specific situational contexts. This aspect is crucial for creating a controlled environment for emotion recognition tasks, where the emphasis is on the tone, pitch, and modulation of the voice rather than the content of the speech itself.

An innovative aspect of CREMA-D is its approach to labelling emotions. Each clip received emotion and intensity labels not just from the actors or a small group of experts, but from a large crowd-sourced platform involving 2,443 raters. This approach provides a broad, democratically sourced perspective on the emotional content of each clip, contributing to the dataset's reliability and applicability across different demographic and cultural groups.

For researchers focusing specifically on audio-based emotion recognition, the audio files extracted from these clips are of primary interest. The audio dimension of CREMA-D includes a wide range of vocal expressions, offering a valuable dataset for training and testing machine learning and deep learning models. The diversity in the actors' performances, coupled with the varied emotional intensities and the comprehensive labelling, makes CREMA-D a powerful tool for advancing research in speech emotion recognition.

By providing a dataset that not only captures a wide spectrum of emotions and intensities but also represents a diverse cross-section of age, gender, and ethnicity, CREMA-D facilitates a more inclusive understanding of emotional expression and recognition in speech. This inclusivity is key for developing more accurate and universally applicable emotion recognition systems.

#### RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), as provided by Livingstone and Russo (2018), stands out in the field of speech emotion recognition due to its detailed and well-structured approach to capturing a wide range of emotional expressions through both speech and song. This dataset's uniqueness lies in its dual modality, offering both audio and visual emotional expressions, enabling researchers to study the impact of multimodal inputs on emotion recognition systems.

With a total of 24 professional actors (12 male and 12 female) contributing to the dataset, RAVDESS provides a diverse and comprehensive collection of emotional expressions. These actors perform two lexically-matched statements across a variety of emotions—calm, happy, sad, angry, fearful, surprise, and disgust—at two levels of intensity, plus a neutral baseline. This range ensures that the dataset covers a broad spectrum of emotional states, facilitating the development of nuanced emotion recognition models.

Each recording in the dataset is meticulously rated by North American participants for three key attributes: emotional validity, intensity, and genuineness. This rigorous evaluation process ensures the reliability and accuracy of the dataset, making it an invaluable resource for training and benchmarking machine learning models in the field of emotion recognition.

The RAVDESS dataset is not only notable for its size, with 7356 recordings, but also for its accessibility and ethical considerations. It is freely available for use in research, provided that it is properly credited, making it an accessible resource for researchers worldwide. Additionally, the creators of RAVDESS have placed a strong emphasis on ethical considerations, including the informed consent of all participants and the respectful treatment of the data, ensuring that the dataset can be used responsibly in research.

#### SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) Database, created by Jackson and Haq (2014), serves as a foundational resource in the domain of speech emotion recognition, a nuanced area of Artificial Intelligence (AI) that focuses on discerning human emotions through speech. This database is specifically designed to facilitate the development and evaluation of automatic emotion recognition systems. It encompasses audio-visual recordings of four male actors, capturing a spectrum of seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral, resulting in a total of 480 British English utterances. These utterances were meticulously selected from the standard TIMIT corpus to ensure a phonetic balance is maintained across the various emotions represented.

The recording process of the SAVEE database was conducted in a visual media lab equipped with state-of-the-art audio-visual recording technologies. This process placed a strong emphasis on capturing phonetically balanced sentences, incorporating a diverse array of emotional expressions to enhance the dataset's effectiveness for emotion recognition tasks. To validate the quality and applicability of the dataset, ten subjects evaluated the recordings under three different conditions: audio, visual, and audio-visual. These evaluations yielded speaker-independent recognition rates of 61%, 65%, and 84%, respectively, highlighting the database's reliability and potential in developing emotion recognition systems.

The SAVEE database includes a broad spectrum of emotions and high-quality recordings, making it a critical tool for researchers in the emotion recognition field. The dataset's specifications feature recordings from four native English male speakers, aged between 27 to 31 years, thereby ensuring a diversity of vocal characteristics. It includes six basic emotions plus a neutral state, encompassing a comprehensive range of human emotional states across 480 utterances. The focus on high recording quality aims to facilitate precise emotion recognition and analysis.

Given its extensive coverage of emotions, meticulous recording quality, and thorough evaluation, the SAVEE database is indispensable for researchers striving to push the boundaries of speech emotion recognition. It supports not only the exploration of audio and visual modalities in isolation but also fosters research into multimodal approaches, which have demonstrated enhanced accuracy in recognizing human emotions. As such, the SAVEE database significantly contributes to advancing AI's capability to interpret human emotions through speech, paving the way for more intuitive and natural human-computer interactions.

#### EMO-DB

The Emo-DB dataset, developed during a DFG-funded research project between 1997 and 1999 by Burkhardt et al. (2005), is a collection of emotional utterances spoken by actors. These recordings were made in the anechoic chamber of the Technical University Berlin, under the guidance of Prof. Dr. W. Sendlmeier, with contributions from Felix Burkhardt, Miriam Kienast, Astrid Paeschke, and Benjamin Weiss. The dataset includes over 500 utterances, categorized by emotions, which are available for research purposes. Users can filter these utterances based on the speaker, text, and emotion through a web interface, which also provides access to syllable labels, duration information, intonation contours, and results from various perception tests.

The structure of Emo-DB allows for detailed analysis of emotional speech, offering tools to analyse fundamental frequency, energy, loudness, duration, stress, and rhythm measurements. The dataset is not only a resource for sound files but also for label files including syllable and phone labels, along with results from perception tests assessing emotion recognition, naturalness evaluation, syllable stress, and emotional intensity. This comprehensive setup aids researchers in understanding how emotional expressions are perceived and processed.

Access to Emo-DB is facilitated through its website, where researchers can download audio and label files for analysis, ensuring the source is correctly cited. The dataset's coding scheme includes a naming convention that integrates the speaker's number, text code, and emotion, providing a systematic approach to data organization. Emo-DB covers a wide range of emotions such as happiness, sadness, anger, fear, and neutral expressions, making it a valuable tool for the study of speech emotion recognition.

The creation and availability of Emo-DB mark a significant contribution to the field of speech emotion recognition, enabling detailed studies of emotional expression in speech. Its structured format, comprehensive content, and accessibility make it a key resource for researchers aiming to analyse and understand emotional expressions in speech, offering a foundation for the development and testing of machine learning and deep learning models in speech emotion recognition.

### Emotions

The diversity and complexity of different and various range of emotions have a significant impact and impose a great obstacle for any model that tries to classify any emotions from an audio dataset. The human complexity of conveying an emotion from diverse ranges of speech pace, linguistics, dialects, and nuances, and the overall combination of them that can be used for similar emotional responses that sound like the direct opposite, seems to be a more improbable mission to dissect and classify it to its true emotion class.

### Authenticity

### Cultural and Gender biases

### Subjectivity and Variability

### Overfitting

NOTE: (PAINDEY ET AL.) The “Conclusion” wraps up by reinforcing the review’s scope, which scrutinized various feature inputs such as Magnitude Spectrogram, Log-Mel Spectrogram, and MFCCs against different architectures to deduce the optimal feature-architecture synergy. Experiments on Emo-DB and IEMOCAP revealed a preference for the Log-Mel Spectrogram feature when paired with the CNN+LSTM architecture. The study acknowledges the challenge of overfitting, especially with limited data as observed with Emo-DB, and details the incorporation of regularization strategies like Dropout and Batch Normalization to mitigate this issue. The article stands firm on the potential of deep learning in SER, underlining the significance of model and feature selection tailored to the complexities of emotional speech data.

# Methodology

## Feature Extraction

## Augmentation

## Model Constructions

### Deep Learning Models

#### ANN

#### CNN

#### MLP

#### ResNet

#### RNN+LSTM

#### Transformer

### Machine Learning Models

#### GMM

#### HMM

#### KNN

#### RF

#### SVM

### Ensemble Models

#### Lorem Ipsum 1

#### Lorem Ipsum 2

#### Lorem Ipsum 3

#### Lorem Ipsum 4

#### Lorem Ipsum 5

## Performance

### Datasets

### Features

## Model Selection

# Own Approach

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Implementation

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Results

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Discussion

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Conclusion

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

**References:**

* + Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access, 9, 47795-47814. doi: 10.1109/ACCESS.2021.3068045.
  + Jaiswal, A., Tiwari, P., Kumar, A., & Tiwari, S. (2020). Speech Emotion Recognition: A Review of Contemporary Techniques. Journal of King Saud University-Computer and Information Sciences. doi: 10.1016/j.jksuci.2020.06.035.
  + P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8462677.
  + W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.
  + S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.
  + C. Zhang and L. Xue, "Autoencoder With Emotion Embedding for Speech Emotion Recognition," in IEEE Access, vol. 9, pp. 51231-51241, 2021, doi: 10.1109/ACCESS.2021.3069818.
  + Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng, Speech emotion recognition: Features and classification models, Digital Signal Processing, Volume 22, Issue 6, 2012, Pages 1154-1160, ISSN 1051-2004, <https://doi.org/10.1016/j.dsp.2012.05.007>.
  + Tin Lay Nwe, Say Wei Foo, Liyanage C De Silva, Speech emotion recognition using hidden Markov models, Speech Communication, Volume 41, Issue 4, 2003, Pages 603-623, ISSN 0167-6393, <https://doi.org/10.1016/S0167-6393(03)00099-2>.
  + Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Trans Affect Comput. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244. PMID: 25653738; PMCID: PMC4313618.
  + Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. <https://doi.org/10.1371/journal.pone.0196391>
  + Jackson, P. and Haq, S., 2014. Surrey audio-visual expressed emotion (SAVEE) database. University of Surrey: Guildford, UK. School of Computer Science and Electronic Engineering. Journal article. Available at: <http://kahlan.eps.surrey.ac.uk/savee/>
  + Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of German emotional speech. In *Interspeech* (Vol. 5, pp. 1517-1520).
  + Fayek, H.M., Lech, M. and Cavedon, L., 2017. Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, pp.60-68.
  + Pandey, S.K., Shekhawat, H.S. and Prasanna, S.M., 2019, April. Deep learning techniques for speech emotion recognition: A review. In 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA) (pp. 1-6). IEEE.

**Acknowledgments:**

* Acknowledge any individuals, organizations, or resources that contributed to your research.

**Appendices:**

* Include any additional material such as code snippets, detailed model architectures, and supplementary information.