SPEECH EMOTION RECOGNITION

Implementing a Supervised Learning Model
CST4050 – Component 3

INTRODUCTION

Emotion, a complex psycho-physiological experience of a person's mental state, has a large impact on human interaction, cognition, perception, and learning. As we incorporate more technology into our daily lives, it becomes increasingly important for robots to recognise, interpret, and respond to human emotions efficiently. This ability to reliably recognise human emotions, specifically from speech, has the potential to transform our interactions with technology, making it more empathic, responsive, and personalized (Zhang et al., 2021). This transformative potential is embodied in the burgeoning field of Speech Emotion Recognition (SER).

SER is a dynamic and challenging Machine Learning (ML) discipline focused on identifying and categorizing emotions embedded in spoken language. The applications of this intriguing technology are vast and profound. In healthcare, SER can facilitate mental health monitoring(Singh et al., 2023), providing objective analysis to assist in the diagnosis and treatment of conditions such as depression, anxiety and autism(Zhang et al., 2021). For customer service, SER can enhance the capabilities of automated voice response systems, enabling them to detect customer frustration or satisfaction and respond appropriately(Morrison et al., 2007). This not only streamlines service but also helps in providing insights for business improvement. Other significant applications are Human-Computer Interaction(HCI)(Cowie et al., 2001), computer games and e-learning systems.

The objective of this project is to develop a robust and reliable SER pipeline with great performance when compared against baselines. The challenge is to design a supervised ML model that accurately recognizes and classifies emotions from speech data. The model should be adept at processing speech data from various speakers, irrespective of their gender, age, or ethnicity, and accurately interpreting a wide range of emotional tones and intensities.

The research question we aim to answer in this project are:

1. Which machine learning models perform best in classifying emotions from the extracted features?
2. How efficient are gradient boosting algorithms in Speech Emotion Recognition task.

To build a comprehensive model that effectively handles the complexities and nuances of human emotion regardless of cultural differences, we leverage three distinct datasets: The Toronto Emotional Speech Set (TESS) provides high-quality emotional speech samples from female actresses (Pichora-Fuller & Dupuis, 2020). The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) offers diverse data from actors of various ages, genders, and ethnic backgrounds (Cao et al., 2014). The Ryerson Audio-Visual Database of Emotional Speech and

Song (RAVDESS) provides emotional speech and song samples from professional actors (Livingstone & Russo, 2018). These datasets enhance our model's generalization, prevent overfitting and help counterbalance gender and cultural bias prevalent in many other datasets used for SER tasks. These datasets are simulated popular resources of speech recognition research created by researchers from various universities and was gotten from their official websites and Kaggle data repository.

REVIEW OF RELEVANT MACHINE LEARNING LITERATURE

Over the past twenty years, emotion recognition has gained significant attention in psychology, biology, and computer science. It has particularly been a focal point in engineering research, aiming to enhance human-computer interaction (HCI) methods. Emotion recognition plays a crucial role in HCI and artificial intelligence (AI), enabling computers to understand and respond to human emotions. Among various modalities, speech signals are a natural carrier of human emotional information and so the Speech Emotion Recognition field has gained the most attention (Zhang et al., 2021).

A lot of research has been done to explore and apply different speech features and techniques in order to extract emotions for emotion recognition. Many classification algorithms, both supervised and unsupervised, have been examined to achieve accurate emotion recognition. However, since emotions are influenced by personal feelings, there is no agreed-upon method for measuring or classifying them. In order to create a successful speech emotion recognition (SER) system, it is important to choose the appropriate speech emotion databases, identify relevant speech features, and develop a versatile classification model (Wani et al., 2021). For this literature review we would explore these factors that determine a successful SER task.

Jahangir et al. (2021) emphasized that hand-engineered feature vectors can pose difficulties when building multilingual SER models due to the unique phonetic and prosodic characteristics of different languages. Recent studies have explored more complex features like Prosodic features related to rhythm, intensity, and stress in speech, and Deep Spectrum features extracted using Convolutional Neural Networks (CNNs) from spectrogram images as demonstrated by Papakostas et al. (2017) on a combination of cross-language publicly available datasets (Emovo, Savee, German).

One of the most commonly utilized features in speech emotion recognition (SER) are Mel Frequency Cepstral Coefficients (MFCCs). These coefficients are highly effective in representing the short-term power spectrum of sound, making them particularly useful for identifying phonetic content. A research paper by Likitha et al. (2017) explores speech emotion recognition using the Mel Frequency Cepstral Coefficient (MFCC) method. It discusses the significance of recognizing emotions from speech signals and its applications in various domains. The paper presents the experimental results, demonstrating an efficiency of approximately 80% in accurately identifying emotions, even in noisy environments.

Another popular feature is the Mel Spectrogram, which provides visual representations of frequency spectra in a sound or other signal as they vary with time. Spectrograms are

advantageous in that they preserve more information about the original audio signal than other feature types. Satt et al. (2017) demonstrated the effectiveness of using spectrograms as input for convolutional recurrent neural networks (CRNNs) models, achieving 68% accuracy in their SER task. The paper introduces a new method for emotion recognition from speech using deep neural networks applied directly to spectrograms. It achieves higher accuracy compared to previous approaches and addresses latency by processing speech in smaller segments.

One last well-known and interesting feature is the Chroma feature, which is a powerful tool for analyzing music due to its ability to represent the harmonic content of audio signals. Chroma features are often used in conjunction with other features, such as MFCCs, to achieve higher accuracy. Krishna et al. (2022) achieved noteworthy accuracy of 86% by utilizing a combination of Chroma, MFCC, and Mel Spectrogram features in Support Vector Machines (SVM) and several other traditional models.

In previous research, SVM and KNN were commonly used. SVM is known for its ability to handle high-dimensional data, while KNN's simplicity and effectiveness make it a popular choice when dealing with irregular decision boundaries. For instance, Jain et al. (2020) utilized SVM using the Linguistic Data Consortium (LDC) Emotional Speech and Transcript database and achieved satisfactory results in their Speech Emotion Recognition(SER) task. Also, Dujaili et al. (2021) proposed a system that extracts speech features like frequency, energy, and Fourier parameters for emotion detection. They used PCA for feature reduction and SVM and KNN for classification. The system showed promising results in accuracy with 88% and execution time using German and English databases.

With the recent surge in deep learning, more complex models have been used, such as CNN, RNN, and DNN. CNNs are popular for their exceptional performance in image and speech recognition, while RNNs, especially the LSTM variant, show potential in SER because of their ability to handle sequential data and retain information for extended periods. Satt et al. (2017) utilised a deep network architecture that combined convolutional and recurrent layers, effectively learning emotional information from spectrograms. The study evaluates different network topologies and parameters, showing improved accuracy for emotion recognition. The experimental results demonstrate the effectiveness of the proposed system in achieving state-of-the-art accuracy and handling noisy environments.

The recent shift towards deep learning models from traditional machine learning models, and the combined use of multiple feature types, has significantly improved the performance of SER systems (Nasim, 2021). However, there are still noticeable gaps in the current research when it comes to the effectiveness of these models across different cultures. The lack of cross-cultural, diverse, and representative datasets hinders the generalizability of the developed SER systems. Additionally, while deep learning models have been quite successful, there has been limited exploration of the potential use of other machine learning models such as gradient boosting decision trees algorithms in the SER domain(Morrison et al., 2007).

This project aims to address these gaps by developing an SER system including gradient boosting decision trees algorithms, trained on a variety of feature types, including MFCCs,

spectrograms, zero cross rate, tonnetz, spectral contrast and chroma features. Furthermore, our SER system will be evaluated on cross-cultural, diverse, and representative datasets.

DESCRIPTION AND JUSTIFICATION FOR THE ML PIPELINE

DATA DESCRIPTION AND PREPROCESSING

The proposed pipeline utilized three separate emotional audio datasets that offer diverse and wide-ranging samples of emotional speech, providing a rich resource for training sophisticated emotion recognition models. The following are the descriptions and features of the three emotional speech datasets, RAVDESS, CREMA-D, and TESS:

1. **Toronto Emotional Speech Set (TESS):**
    - **Dataset Size:** 2800 audio files
    - **Source:** Two female actresses aged 26 and 64 years.
    - **Emotions:** Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, and Neutral.
    - **Special Features:** Emphasizes female speakers, counterbalancing the male bias in many other datasets.
2. **Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D):**
    - **Dataset Size:** 7,442 original clips
    - **Source:** 91 actors (48 males and 43 females) aged 20 to 74, representing diverse racial and ethnic backgrounds such as African American, Asian, Caucasian, Hispanic, and others not specified.
    - **Emotions:** Anger, Disgust, Fear, Happiness, Neutral, and Sadness.
    - **Intensity Levels:** Low, Medium, High, or unspecified level.
    - **Special Features:** Variety in actors and their backgrounds. Also includes varied intensity levels for each emotion.
3. **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):**
    - **Dataset Size:** 1440 audio files
    - **Source:** 24 professional actors (12 female, 12 male), delivering phrases with a neutral North American accent.
    - **Emotions:** Calm, Happiness, Sadness, Anger, Fear, Surprise, Disgust, and Neutral.
    - **Intensity Levels:** Normal and Strong (alongside an additional Neutral expression).
    - **Special Features:** Inclusion of both speech and song samples, variety in emotion intensity levels.

The datasets were loaded using their audio file paths and the emotional labels were extracted from the filenames. For RAVDESS, numerical codes are mapped to corresponding emotion names, while CREMA uses three-letter codes and TESS uses actual emotion names. These extracted paths and labels are subsequently organized into pandas DataFrames, each featuring two columns: 'Emotion' and 'Path'. These represent the emotional labels and the full file paths to the audio files, respectively. Signal processing was performed on the audio

file paths and data augmentation techniques such as noise addition, time-stretching and pitching were carried out on the extracted audio signals so that the model could generalize better and prevent overfitting as our dataset are simulated.

FEATURE EXTRACTION

Extraction of various acoustic features from the audio signals were done. The extracted features include:

1. Mel-frequency cepstral coefficients (MFCCs): MFCCs capture the spectral characteristics of the speech signal and are commonly used in speech-related tasks.
2. Chroma feature: Chroma represents the distribution of musical pitch classes in the audio signal and can provide information about the tonal content of speech.
3. Spectral contrast: Spectral contrast measures the difference in magnitudes between peaks and valleys in the spectrum and can capture the spectral richness and brightness of the speech signal.
4. Tonnetz: Tonnetz is a harmonic-based feature that represents the harmonic relationships between different musical tones and can capture tonal properties in the speech.
5. Zero-crossing rate (ZCR): ZCR measures the rate at which the audio waveform changes its sign and can provide information about the temporal characteristics of the speech signal.

These features were selected because they have been found to be relevant and informative for speech emotion recognition tasks.. By combining these diverse features, the feature extraction process aims to capture various aspects of the speech signal that are indicative of emotional content.

After feature extraction, the features were flattened, meaning that the multidimensional feature matrices were converted into one-dimensional vectors. This flattening step is often performed before feeding the features into a machine learning model for classification and this process ensures compatibility with a wide range of machine learning algorithms and libraries.

TRAIN-TEST SPLIT

After feature extraction, the final dataframe, which was made up of 22468 rows and 388 columns, was split into target and predictors dataframe, after which label encoding was performed on the target dataframe for traditional models and deep learning models, one-hot encoding was separately done. The dataframes were then split into Train, Validation and Test sets for traditional models and deep learning models respectively in order to evaluate model performance, detect overfitting, tune hyperparameters, compare models, and ensure the integrity of the data. with the target variable ‘y’ stratified due to the imbalance in the dataset so as to ensure balanced representation of classes in all sets.

BASELINE MODELS

1. Decision Trees: Decision Trees are intuitive and easy to interpret models. They make decisions by recursively partitioning the feature space based on feature values.

Decision Trees can handle both categorical and numerical features, making them suitable for speech emotion recognition tasks where the input features can be diverse and imbalanced as in the case of our dataset. Decision Trees are also robust to outliers and can capture nonlinear relationships between features and emotions. They serve as a good starting point for understanding the importance of different features in the classification task. that can handle imbalanced data

2. Naive Bayes: Naive Bayes classifiers have been successfully applied to text classification tasks, and speech emotion recognition can be framed as a similar classification problem. Naive Bayes is known for its simplicity, efficiency, and ability to handle large feature spaces, making it a good choice as a baseline model, however its parameters would be adjusted to handle the imbalance in the dataset.

3. Logistic Regression: Logistic Regression is a widely used linear classification model. Logistic Regression is interpretable and can provide insights into the influence of individual features on the classification decision. It can handle both binary and multiclass classification problems, making it suitable for speech emotion recognition with multiple emotion classes, but would also require class weight adjustment to perform well on imbalanced data.

FEATURE SELECTION

Feature Importance: Random Forest ensemble algorithm was employed for getting the importance of each feature. The larger score indicated higher importance. The features were selected by dropping features that were lowere that the set threshold of 0.02.

ADVANCED MODELS

Gradient Boosting Classifiers:

1. **XGBoost:** is an effective choice for Speech Emotion Recognition (SER) due to its ability to handle mixed data types, perform automatic feature selection, resist overfitting through regularization, operate with high efficiency and speed, and work with a variety of problem types including multi-class classification. Moreover, it provides interpretable feature importance scores, offering insights into which features are most influential in predictions.
2. **CatBoost:** It handles categorical variables effectively, is robust to noisy data, and automatically scales features and handles missing values. CatBoost's excellent performance, efficient training capabilities, and interpretability make it a strong choice for SER tasks.

Traditional Models:

1. **SVM:** VM is a powerful algorithm for classification tasks, including SER. It works by finding an optimal hyperplane that maximally separates different classes of data. SVM is particularly useful when dealing with high-dimensional feature spaces and can handle both linear and non-linear relationships between features. It has good generalization capabilities and can handle small to medium-sized datasets. SVM also

supports the use of different kernel functions, such as linear, polynomial, and radial basis function (RBF), allowing it to capture complex relationships in the data.

2. **KNN:** KNN is a simple yet effective algorithm for classification tasks, including SER. It classifies new instances based on the majority vote of the K nearest neighbors in the feature space. KNN is a non-parametric algorithm, meaning it does not make strong assumptions about the underlying data distribution. It is relatively easy to implement and can handle multi-class classification problems. KNN is robust to noisy data and can capture local patterns in the feature space. It can also be adapted to handle dynamic or evolving datasets

Deep Learning Model:

CRNN: The CRNN model is effective for Speech Emotion Recognition due to its ability to extract relevant features, capture sequential information, learn hierarchical representations, handle variability, and perform end-to-end learning. It combines CNNs and RNNs for feature extraction, temporal modeling, and emotion classification, making it a comprehensive and robust solution for SER tasks.

HYPERPARAMETER TUNING:

This is carried using Bayesian Optimization. The following were how the parameters were tuned

1. XGBoost hyperparameters: {'colsample_bytree': 0.5478318452218058, 'gamma': 0.048627005289563874, 'learning_rate': 0.09750214749532385, 'max_depth': 18.0, 'min_child_weight': 1.5465616685882315, 'n_estimators': 169.0, 'subsample': 0.7420530511989842}
2. SVM Hyperparameters: {'C': 6.579861091236223, 'gamma': -0.3581757594907211, 'kernel': 'rgb'}
3. KNN Hyperparameters:{'n_neighbors': 14.0, 'p': 0, 'weights': 'uniform'}
4. CRNN: {'batch_size': 0, 'dropout': 0.05202236068559801, 'epochs': 2, 'optimizer': 1, 'units': 1}

METRICS

Classification report and confusion matrix were the evaluation metrics for this speech emotion recognition (SER) tasks with because of the imbalanced multiclasses.

1. Classification Report:

- Precision, Recall, and F1-score: In imbalanced datasets, the distribution of classes is uneven, and accuracy alone may not provide a complete picture of model performance. The classification report provides precision, recall, and F1-score metrics for each class, which are more informative. These metrics consider true positives, false positives, and false negatives, providing insights into how well the model performs for each emotion class individually. This is particularly important when dealing with imbalanced classes, as it helps identify whether the model is biased towards the majority class or performs well across all classes.

2. Confusion Matrix: The confusion matrix is a helpful tool for visualizing a model's performance in classifying emotions. It can also identify class imbalance issues and guide improvements in the model or data collection process.

Overall, the combination of classification report and confusion matrix provided a comprehensive evaluation of the model's performance in SER tasks with imbalanced multiclasses. They help in understanding the model's behavior for each emotion class, detecting biases, and identifying class imbalance issues, leading to insights for further improvements in the model and data collection strategies.

MODEL EVALUATION

Baseline Models:

Decision Tree: The model has good precision and recall for most classes, with class 1 and 7 having the highest values. It achieves a balanced F1-score for most classes and an overall accuracy of 0.71.The model achieves an overall accuracy of 0.71, indicating the proportion of correctly predicted instances out of the total instances.

Logistic Regression: The model has varying precision and recall values for different classes. Overall accuracy is 0.63. Class 1 has the highest precision, recall, and F1-score. Some classes have relatively low precision values, indicating a higher rate of false positives.

Naive Bayes: The model has varying precision and recall values for different classes, with class 1 and 7 having the highest precision and class 1 and 5 having the highest recall. The F1-scores are relatively low for most classes, with class 1 having the highest score. The overall accuracy of the model is 0.27.

Traditional Models:

XGBoost: Precision, recall, and F1-scores: XGBoost achieves high performance across most classes, with balanced precision, recall, and F1-scores. The model demonstrates good accuracy in predicting positive instances for each class, resulting in high scores. The model achieves an overall accuracy of 0.85, indicating the proportion of correctly predicted instances out of the total instances.

CatBoost: Precision, recall, and F1-scores: CatBoost exhibits good performance with balanced precision, recall, and F1-scores for most classes. The model accurately predicts positive instances for each class. The model achieves an overall accuracy of 0.82, indicating the proportion of correctly predicted instances out of the total instances.

SVM: Precision, recall, and F1-scores: SVM demonstrates reasonable precision, recall, and F1-scores for most classes. The model performs well in identifying positive instances. The model achieves an overall accuracy of 0.77, indicating the proportion of correctly predicted instances out of the total instances.

KNN: Precision, recall, and F1-scores: KNN achieves relatively balanced precision, recall, and F1-scores for most classes. The model demonstrates good accuracy in predicting positive

instances. The KNN model achieves an overall accuracy of 0.68, indicating the proportion of correctly predicted instances out of the total instances.

EVALUATION METRICS COMPARISON FOR SER MODEL

MODELS	TRAIN	CONFUSION MATRIX	ACCURACY	F1 SCORE
DECISION TREE	0.71	0.75	0.73	0.71
LOGISTIC REGRESSION	0.60	0.64	0.63	0.66
NAIVE BAYES	0.22	0.31	0.27	0.26
XGBOOST:	0.80	0.98	0.85	0.87
CATBOOST:	0.79	0.96	0.82	0.85
SVM	0.59	0.87	0.77	0.80
KNN	0.66	0.73	0.68	0.71
CRNN	0.65	0.75	0.77	0.73

Decision Tree Classification report:

Decision Tree Performance:

	precision	recall	f1-score	support
0	0.77	0.75	0.76	895
1	0.78	0.87	0.82	92
2	0.68	0.65	0.67	894
3	0.67	0.69	0.68	894
4	0.67	0.67	0.67	895
5	0.72	0.72	0.72	760
6	0.71	0.72	0.71	894
7	0.78	0.82	0.80	284
accuracy			0.71	5608
macro avg	0.72	0.74	0.73	5608
weighted avg	0.71	0.71	0.71	5608

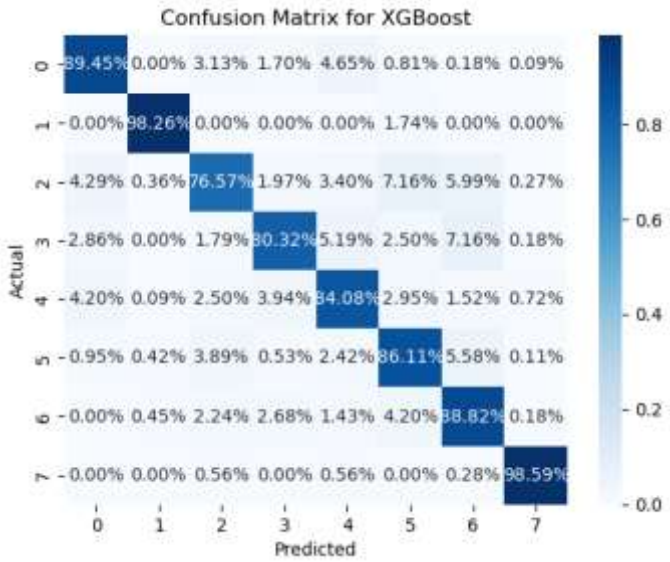
XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.88	0.89	0.89	1118
1	0.89	0.98	0.93	115
2	0.85	0.77	0.81	1118
3	0.88	0.80	0.84	1118
4	0.83	0.84	0.84	1118
5	0.80	0.86	0.83	950

```
              6       0.82      0.89      0.85      1118
              7       0.95      0.99      0.97       355

    accuracy                            0.85      7010
   macro avg        0.86      0.88      0.87      7010
weighted avg        0.85      0.85      0.85      7010
```

XGBoost Confusion Matrix Heatmap:



XGBoost Actual vs Predicted Classes:

```
       Actual Predicted
0     disgust   disgust
1       angry     angry
2         sad       sad
3       happy     happy
4         sad       sad
...       ...       ...
7005     fear      fear
7006    happy     happy
7007    happy     happy
7008  neutral   neutral
7009     fear      fear
```

Comparing Baseline Models and Complex Models:

From the above comparison, it can be observed that the traditional models (XGBoost, CatBoost, SVM, and KNN) generally outperform the baseline models (Decision Tree, Logistic Regression, and Naive Bayes) in terms of accuracy and F1-Score. The CRNN model performs comparably to the traditional models, with a similar accuracy and slightly lower F1-Score. They exhibit better performance in predicting positive instances for most classes, resulting in higher overall scores and also demonstrate the ability to handle imbalanced classes more effectively, achieving better performance across all classes.

CONCLUSION

In summary, the baseline models (Decision Tree, Logistic Regression, and Naive Bayes) show varying performance on the speech emotion recognition task. The Decision Tree model demonstrates decent precision and recall, while Logistic Regression struggles with imbalanced classes but performs well in terms of F1-score. Naive Bayes, however, exhibits limitations in handling imbalanced data. Comparatively, the traditional models (XGBoost, CatBoost, SVM, and KNN) generally outperform the baselines, with higher accuracy and F1-scores. Among them, XGBoost and CatBoost achieve notable precision, recall, and F1-scores. Finally, the CRNN model shows promising results with high accuracy, F1-score, and test performance metrics. Limitations include imbalanced classes and performance discrepancies, suggesting the need for addressing class imbalance, exploring alternative features, and utilizing ensemble methods, hyperparameter tuning, and data augmentation techniques. Furthermore, recommendations include leveraging deep learning architectures, transfer learning, dataset expansion, and model interpretability techniques. These strategies can enhance the performance and robustness of the models in speech emotion recognition tasks.

REFERENCES

1. Zhang, S., Liu, R., Tao, X., & Zhao, X. (2021). Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives. *Frontiers in Neurorobotics*, *15*. https://doi.org/10.3389/fnbot.2021.784514
2. Wani, T. M., Gunawan, T. S., Qadri, S. B., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, *9*, 47795–47814. https://doi.org/10.1109/access.2021.3068045
3. Ayadi, M. M. H. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020
4. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, *21*(4), 1249. https://doi.org/10.3390/s21041249
5. Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal*

of Speech Technology, *21*(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

6. Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., & Makedon, F. (2017). Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition. *Computation*, *5*(4), 26. https://doi.org/10.3390/computation5020026
7. Jahangir, R., Wah, T. Y., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*, *80*(16), 23745–23812. https://doi.org/10.1007/s11042-020-09874-7
8. Likitha, M. S., Gupta, S. C., Hasitha, K., & Raju, A. U. (2017). *Speech based human emotion recognition using MFCC*. https://doi.org/10.1109/wispnet.2017.8300161
9. Satt, A., Rozenberg, S., & Hoory, R. (2017). *Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms*. https://doi.org/10.21437/interspeech.2017-200
10. Krishna, K. V., Sainath, N., & Posonia, A. M. (2022). Speech Emotion Recognition using Machine Learning. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*. https://doi.org/10.1109/iccmc53470.2022.9753976
11. Jain, M., Narayan, S., Balaji, P., & Muthu, R. (2020). Speech Emotion Recognition using Support Vector Machine. *ResearchGate*. https://www.researchgate.net/publication/339350511_Speech_Emotion_Recognition_using_Support_Vector_Machine

12. Nasim, S. (2021, September 2). *Speech Emotion Recognition Using Deep Learning*. https://blog.dataiku.com/speech-emotion-recognition-deep-learning
13. Singh, J., Babu-Saheer, L., & Faust, O. (2023). Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health*, *20*(6), 5140. https://doi.org/10.3390/ijerph20065140
14. Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, *49*(2), 98–112. https://doi.org/10.1016/j.specom.2006.11.004
15. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G. N., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80. https://doi.org/10.1109/79.911197
16. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
17. Cao, H., Cooper, D. A., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. https://doi.org/10.1109/taffc.2014.2336244
18. Dujaili, M. J. A., Ebrahimi-Moghadam, A., & Fatlawi, A. H. (2021). Speech emotion recognition based on SVM and KNN classifications fusion. *International Journal of Power Electronics and Drive Systems*, *11*(2), 1259. https://doi.org/10.11591/ijece.v11i2.pp1259-1264

19. Pichora-Fuller, M. K., & Dupuis, K. (2020). *Toronto emotional speech set (TESS)* [Dataset]. https://doi.org/10.5683/sp2/e8h2mf