Audio Data: Speech Emotion Recognition

1. Introduction

The purpose of the communication is to exchange information, inform people about the incidents, influence people to do work, express ideas, etc. Emotion plays an imperative role during communication. It helps people to build trust, show empathy, influence, and win in negotiation. When a person communicates with another person, emotion gives insight into what action she or he should take.

Even though it is easy for a human to understand the emotions like sad, happy, neutral, angry etc., it is challenging for the machine to extract emotion (El Ayadi et al., 2011, Schuller et al., 2011). AI (Artificial Intelligence) has made great progress in speech recognition, text to speech, speech to text, and sentiment analysis. It is still far off to understand emotions and interact with humans (Han et al., 2014, Ling et al., 2015).

Speech emotion recognition could be one of the steps for the machine to understand our emotion and react. The application of this would-be automatic identification of customer satisfaction in a call centre, an ambient system that reacts based on mood, solves various language ambiguities, etc.

Furthermore, speech emotion recognition is one of the focal research areas in the past few years (Tripathi et al., 2011). Therefore, this paper aims to explore machine learning model that detects emotion from the speech. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset will be used in this project and main is to identify emotion from the audio which has higher value based on MFCC (Livingstone and Russo, 2018, Logan, 2000).

2. Literature Review

Supervised learning consists of input-output pairs; it learns a function from labelled training data. There are different supervised learning algorithms, including linear regression, logistic regression, decision tree, random forest etc. (Caruana and Niculescu-Mizil, 2006). The various supervised algorithm uses to identify emotion from speech (Koolagudi and Krothapalli, 2012). The below table provides a summary of previous work on emotion recognition from speech.

Table 1: Summary of previous work on emotion recognition from speech

Reference	Model	Database	Accuracy	Notes
Iqbal and Barua (2019)	Gradient Boosting, KNN (K-Nearest Neighbor) and SVM (Support Vector Machine)	RAVDES	80 %	They develop a real-time emotion recognition system by analysing tonal properties which can detect four emotions anger, happiness, sadness and neutral. They use several features including MFCC, Pitch etc. Gradient boosting performs better in all category in compare with KNN and SVN.
Sinith et al. (2015)	SVM	Berlin emotional database	75 %	They use different combination of features including MFCC, Pitch and energy.
Jannat et al. (2018)	CNN (Conventional Neural Network)	RAVDES	66.41% in audio data, 90% in audio + video data	They use a combination of audio and video data. It got better accuracy when combining audio and video data but got less accuracy when using audio data only.
Huang and Bao (2019)	SVM and CNN	RAVDES	CNN 85% SVM 48.11%	They extract MFCC (spectrum of-a-spectrum) features from audio files and classify seven emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised).

3. Problem Statement

Identify emotion (ex. sad, happy, angry, neutral) of the speaker from the given voice input.

The input to the model:

- Speech sample from any speaker.

Output:

- Emotion of the speaker.

4. Dataset

The dataset is built using 1440 samples from RAVDES of Speech file, and this dataset includes recordings of 24 professional actors (12 females, 12 males), which provides for calm, happy, sad, angry, fearful, surprise, and disgust expressions (Livingstone and Russo, 2018). All the audio files are 16 bit with .wav format.

All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav)

Data visualization

Use waveplot to plot a signals amplitude envelope over time.

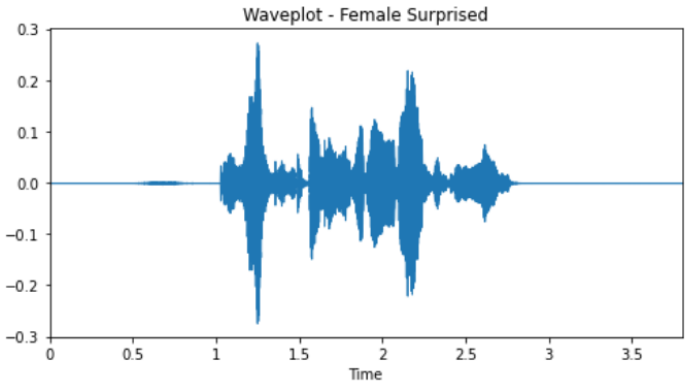

Figure1: Sample audio, Female actor in surprised emotion.

Table 2: Number of emotions and gender distribution of dataset.

Emotion	Fearful	Angry	Disgust	Happy	Surprised	Calm	Neutral	Sad
	192	192	192	192	192	192	96	192
Gender	Male	Female						
	720	720						

The dataset has equal number of male and female actors sample audios. It has also equal number (192) of emotions file except neutral emotion (96).

5. Benchmark

Huang and Bao (2019) used SVM and got 48.11% accuracy. They benchmarked model with following parameters SVM (kernel='linear' and 'RBF', c =10, gamma = Scale). The MFCC feature range from 10 to 120. There is a positive correlation between the MFCC and linear kernel model (Huang and Bao, 2019), and it's a popular model for classification (Campbell et al., 2006). With 100 MFCCs, they can achieve

better accuracy (Huang and Bao, 2019). There, SVM will be used as a baseline test, where MFCC = 100 and Kernal=linear)

6. Evaluation Metric

Recall, precision, f1-score, and confusion matrix are poplar metric for the audio emotion recognition task (Zhang et al., 2016; Huang and Bao, 2019; Rintala, 2020). In this project, I have used the same metric to evaluate the performance of the model. Furthermore, it's an imbalance data, precision and recall work lll.

The F1 score combines both precision and recall and gives harmonic value. It reaches its best value as it tends to 1 (Goutte and Gaussier, 2005). I will compare the scores across our experiments to present conclusive evidence on which method gives the best results for detecting speech from audio samples. Furthermore, there may be a use case to extract audio emotion in real-time, in that case, time to run the model is also essential, along with accuracy (Stuart et al., 2020). Therefore, I have considered time as a metric in this project.

K-fold Cross-validation technique will be considered to check the model ability to predict new data. In cross-validation, low variance means low bias; hence model has a higher chance of predicting new data (Bengio, and Grandvalet, 2004).

Confusion metrics provides performance of classification model (Jayaswal, 2020).

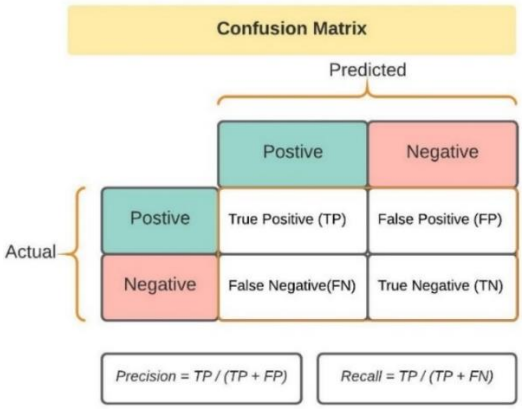

Figure 2: Confusion matrix

- TP (Model correctly predicts the positive class)
- TN (Model correctly predicts the negative class)
- FP (Model gives the wrong prediction of the negative class)
- FN (Model wrongly predicts the positive class)

F1 Score takes both false positive and false negatives into account.

$$F1\ Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

7. Methodology

I have used python libraries Librosa, Soundfile, Sklearn etc. to test different model SVM, Decision Tree, Random forest, which can recognize emotion from an audio file. The first step is to load the dataset, extract the features and store it in a panda's data frame. Then split the dataset with training and testing sets. Finally use a different machine learning model to train the data and test it and find out the best model based on F1 score and time.

Mfcc: Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound (Iqbal and Barua, 2019), is extracted from the audio file. I use this because of its low complexity and better ability to extract the feature from speech. (Zeng et al., 2008).

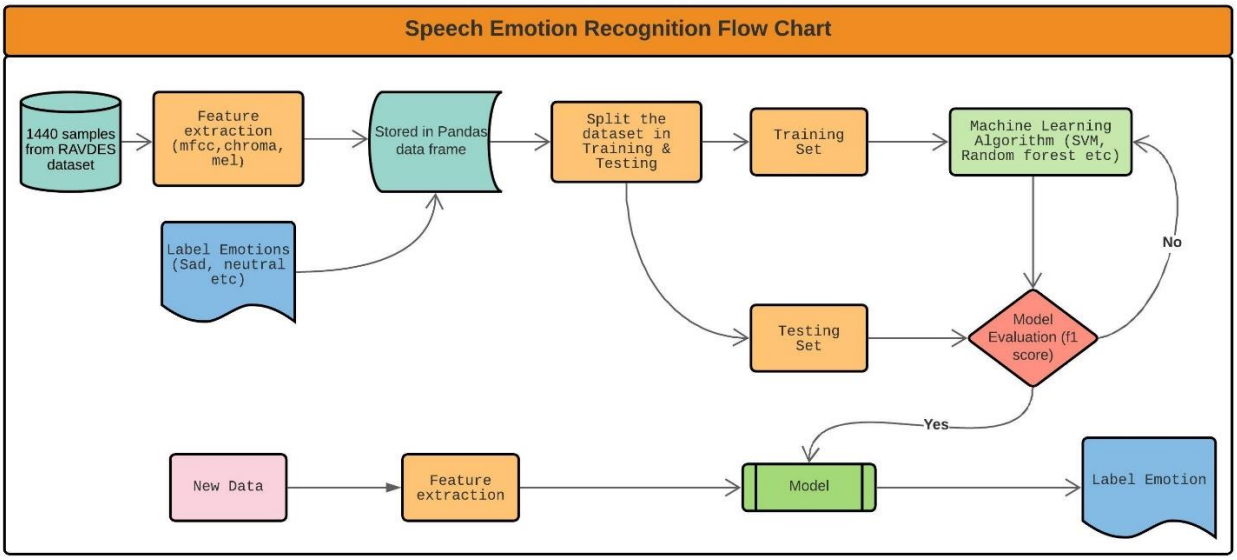

Figure 3: Block diagram of speech emotion recognition

Among these features, MFCC is a dominant factor, and it provides rich feature content (Huang and Bao, 2019). Therefore, I have used MFCC as the main feature for this project. There is a class imbalance in neutral emotion (96) with other emotions 192. However, Huang and Bao (2019) found that treating the class imbalance problem did not improve model performance.

8. Classification Algorithms

Tarunika et al., (2018), Iqbal and Barua (2019) use KNN in their paper. Other researchers use SVM (Iqbal and Barua, 2019; Sinith et al. 2015; Huang and Bao, 2019), decision tree (Lee et al., 2004). In this project, I have used SVM, KNN and Random forest. Instead of the decision tree, I have used the random forest as decision tree tends to overfit the data (Ali et al., 2012). They have all been implemented in python using specialized libraries that offer parameter optimization and evaluation metrics.

SVM (Support Vector Machine)

It's advantageous where there is a high dimensionality in the dataset. It has a different kernel which allows us to convert any linear model to non-linear one to achieve a better result (Iqbal and Barua, 2019)

SVM tries to find the optimal line that maximally separates the two classes, which can be used for multi classes.

$$g(x) = w^T x + b = 0$$

Based on the value of g for a particular point x, its possible to find out two classes. A similar concept can be used for multiclass. At the heart of the model is the kernel function. Scikit-learn package has been used to create an SVM model where the model's main parameter is Kernal function, which takes the given input and transform it into the required form.

KNN (K Nearest Neighbour)

It classifies features using a nearest neighbour interpolation method. It captures the idea of similarity; we can change the K value to determine the better accuracy where K means the number of the nearest neighbourhood (Iqbal and Barua, 2019). KNN calculates the distance between data points

$$d\,(x,y) = d\,(y,x) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 \dots\dots\dots\dots\dots + (x_n - y_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_n - y_n)^2}$$

Random Forest

Random forest uses multiple decision trees at training time and outputting the class that is the class model in classification (Ali et al., 2012. The key parameters here are the number of trees, number of samples and number of features.

Table 3: Advantages and disadvantages of different algorithms

Algorithms	Advantages	Disadvantages
SVM (Support Vector Machine)	Memory efficient Works well when there is a clear separation margin	Not suitable for large dataset Not performing well when there is lots of noise in dataset.
KNN	Easy implementation No training period so new data can be added any time	Not suitable for large dataset Sensitive to noisy and missing data
Random Forest	Reduce overfitting of data Less impacted by nosie	Complex model Time consuming

9. Training and Hyperparameter Tuning:

The below table shows different algorithms which have used in this project and their hyperparameters.

Table 4: Algorithms and hyperparameter

Algorithms	Hyperparameter
SVM (Support Vector Machine)	Kernal = "linear", "rbf", "poly", "sigmoid", it's a main hyperparameter in SVM C is an error control hyperparameter in SVM, low C means less error and high C means, large error. The gamma parameter (g) value defines how far the influence of a single training example reaches, with low value means "distant" and high value meaning "near." (sklearn.svm.SVR — scikit-learn 0.24.0 documentation, 2020)
KNN	It classifies new data based on its closeness to K-neighbors. K = 5 means 5 training data closets to the new data. Hyperparameter p, p = 1 means Manhattan distance and p = 2 means Euclidean distance. (Adipta, 2020)
Random Forest	n_estimators - number of trees in the forest (default - 100) max_features - auto;(default is auto, i.e., number of features considered while developing a tree max_samples - How many samples should be drawn from the dataset to train each base estimator (Koehrsen, 2020)

I have tried popular train-test split 80%-20%, 67%-33%, 50%-50% and chosen 67%-33% based on the accuracy.

The hyperparameter of the model has been optimized with "GridSearchCV" technique. This library has been imported from Sklearn (sklearn.model_selection.GridSearchCV — scikit-learn 0.24.0 documentation, 2020).

10. Results

Without tunning any hyperparameters, Random forest performed better avg accuracy 55.00% followed by SVM avg accuracy 49.00 % with MFCC = 100. The accuracy is slightly higher than the benchmark 48.11% accuracy, which was achieved in SVM.

Table 5: Performance of the model without tunning hyperparameters

Emotions	SVM (f1 Score)	KNN (f1 Score)	Random forest (f1 score)
Angry	0.62	0.60	0.57
Calm	0.54	0.59	0.66
Disgust	0.41	0.35	0.52
Fearful	0.53	0.43	0.59
Happy	0.47	0.35	0.47
Neutral	0.39	0.28	0.33
Sad	0.46	0.35	0.47
Surprised	0.48	0.39	0.62
Avg f1	0.49	0.42	0.53

Table 6: Training vs test accuracy of algorithms

Algorithms	Training accuracy (%)	Test accuracy (%)
SVM (Support Vector Machine)	69.61	55.25
KNN	65.25	55.25
Random Forest	100.00	55.25

Although Random forest achieved high accuracy, there is a high difference between training and test accuracy, which cause biases. Comparing the bias, SVM performs better among the three models without hyperparameter.

After tunning, the hyperparameter, KNN achieved high f1 score 58.00 % where leaf_size = 1, n_neighbors =2, p =1 (Manhattan distance). The value is much higher than the benchmark of 48.11%.

Furthermore, five-folds cross-validation has been run where KNN has less variance compared to SVM, and it takes less time (0.02 ms) to train the model.

Table 7: Performance of the model with tunning hyperparameters

Emotions	SVM (f1 Score) Hyperparameters C=100, gamma=0.001, Kernal=rbf	KNN (f1 Score) Hyperparameters leaf_size = 1, n_neighbors =2, p =1	Random forest (f1 score) Hyperparameters n_estimators = 250, max_samples=0.8, max_features=0.25, random_state=42
Angry	0.62	0.69	0.62
Calm	0.54	0.74	0.65
Disgust	0.41	0.63	0.52

Fearful	0.53	0.67	0.66
Happy	0.47	0.55	0.50
Neutral	0.39	0.58	0.34
Sad	0.46	0.62	0.44
Surprised	0.48	0.61	0.62
Avg f1	0.56	0.58	0.57
Cross_Validation_ Variance	0.05	0.02	0.02
Time train the model (ms)	0.19	0.02	2.54

11. Conclusion

I have successfully detected emotion from audio samples and compared three models based on f1-score, K-fold cross-validation and time. To summarize, KNN performed better in compared with SVM and Random forest. Even though KNN achieved greater accuracy of 58.00%, still need to improve accuracy.

The biggest challenges of this project were feature extraction, hyper tune and found the best model. As initially, SVM performed better; however, all the models performed better after training the model with selected hypermeters. K-fold cross-validation technique helped to find the best model.

Audio emotion recognition is an emerging field. It can be used in a different use case. In the future, we can explore the deep learning model to see how accuracy varies. Furthermore, it can be possible to try MFCC with other features like Chroma, Mel.

12. Reference

Adipta Martulandi, M., 2020. K-Nearest Neighbors In Python + Hyperparameters Tuning. [online] Medium. Available at: <https://medium.com/datadriveninvestor/k-nearest-neighbors-in-python-hyperparameters-tuning-716734bc557f> [Accessed 25 December 2020].

Ali, J., Khan, R., Ahmad, N. and Maqsood, I., 2012. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), p.272.

Bengio, Y. and Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research, 5(Sep), pp.1089-1105.

Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E. and Torres-Carrasquillo, P.A., 2006. Support vector machines for speaker and language recognition. Computer Speech & Language, 20(2-3), pp.210-229.

Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).

El Ayadi, M., Kamel, M.S. and Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), pp.572-587.

Goutte, C. and Gaussier, E., 2005, March. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European conference on information retrieval (pp. 345-359). Springer, Berlin, Heidelberg.

Han, K., Yu, D. and Tashev, I., 2014. Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association.

Huang, A. and Bao, P., 2019. Human vocal sentiment analysis. arXiv preprint arXiv:1905.08632.

Koolagudi, S.G. and Krothapalli, S.R., 2012. Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. International Journal of Speech Technology, 15(4), pp.495-511.

Iqbal, A. and Barua, K., 2019, February. A Real-time Emotion Recognition from Speech using Gradient Boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

Jannat, R., Tynes, I., Lime, L.L., Adorno, J. and Canavan, S., 2018, October. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Iarable Computers (pp. 956-959).

Jayaswal, V., 2020. Performance Metrics: Confusion Matrix, Precision, Recall, And F1 Score. [online] Medium. Available at: <https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262> [Accessed 22 December 2020].

Koehrsen, W., 2020. Hyperparameter Tuning The Random Forest In Python. [online] Medium. Available at: <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74> [Accessed 25 December 2020].

Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S., 2004. Emotion recognition based on phoneme classes. In Eighth International Conference on Spoken Language Processing.

Ling, Z.H., Kang, S.Y., Zen, H., Senior, A., Schuster, M., Qian, X.J., Meng, H.M. and Deng, L., 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. IEEE Signal Processing Magazine, 32(3), pp.35-52.

Livingstone, S. and Russo, F., 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), p.e0196391.

Logan, B., 2000, October. Mel frequency cepstral coefficients for music modeling. In Ismir (Vol. 270, pp. 1-11).

Tripathi, S., Kumar, A., Ramesh, A., Singh, C. and Yenigalla, P., 2019. Deep learning-based emotion recognition system using speech features and transcriptions. arXiv preprint arXiv:1906.05681.

Rintala, J., 2020. Speech Emotion Recognition from Raw Audio using Deep Learning.

Schuller, B., Batliner, A., Steidl, S. and Seppi, D., 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9-10), pp.1062-1087.

Scikit-learn.org. 2020. Sklearn.Svm.SVR — Scikit-Learn 0.24.0 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html> [Accessed 26 December 2020].

Scikit-learn.org. 2020. Sklearn.Model_Selection.Gridsearchcv — Scikit-Learn 0.24.0 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html> [Accessed 26 December 2020].

Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P. and Rajan, S., 2015, December. Emotion recognition from audio signals using Support Vector Machine. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 139-144). IEEE.

Stuart, C., Harrison, R., Jonathan, W. and Rich, P., 2020. Supervised machine learning for audio emotion recognition. Personal and Ubiquitous Computing.

Tarunika, K., Pradeeba, R.B. and Aruna, P., 2018, July. Applying machine learning techniques for speech emotion recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Zhang, B., Provost, E.M. and Essl, G., 2016, March. Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5805-5809). IEEE.

Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S., 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), pp.39-58.