Grenze International Journal of Engineering and Technology, June Issue

Speech Emotion Recognition using Machine Learning: A Review

Ama Gyekyewaa Odei-Tettey[1] and Govindha Ramaiah Yeluripati[2]
[1-2]Computer Science Department, Lancaster University Ghana, Accra, Ghana
Email: {amaodeitettey, g.yeluripati}@gmail.com

***Abstract*—Affective computing, the area of technology which the topic of this research paper falls under, is concerned with sensing the emotions of users and using the data gathered from the inputs that these users provide to serve them better. Speech Emotion Recognition, or SER for short, has a growing number of applications, including intelligent toys and psychiatric diagnoses. Despite the strides that have been made in this area of computing from its inception in 1996, there is much room for improvement and if these improvements are made, SER could prove useful to recreational users and professionals across industries. This paper aims to comparatively analyze various machine learning approaches to emotion recognition using speech, and to attempt to suggest optimal techniques for use throughout the development process to support the realization of this goal.**

***Index Terms*— Speech Emotion Recognition, Affective Computing, Deep Neural Networks, Machine Learning, Spectral Features, Prosodic Features, Dataset Creation, Linear Classifier, Non-linear Classifier.**

I. INTRODUCTION

Affective computing systems can be categorised into 4 main groups: Speech Emotion Recognition, Facial Emotion Recognition, Textual Emotion Recognition and Physiological Signals [8]. Speech Emotion Recognition, in particular, since its advent in 1996 [9], has shown much promise in enhancing Human-Computer Interaction. Out of the aforementioned affective computing categories, SER is considered to be the easiest and most straightforward means of emotion data collection and processing [2]. For this reason, it is applied in technology used across industries. Some of these applications include interactive tutorials and courses, psychotherapy and counselling appointments and ensuring vehicle safety for motorists [1].

Traditionally, SER systems have been constructed with classical machine learning algorithms such as the Hidden Markov Model and Gaussian Mixture Model. While algorithms like these are known to ensure some level of accuracy and have contributed to the many strides that SER has made over the years, it is worth investigating whether newer machine learning technology can augment what is already considered to be a remarkable accomplishment in the area of Speech Emotion Recognition. In addition to the above, it is also worth evaluating common practices involved in the steps to realising an SER system. This includes dataset creation or selection, which refers to how samples are collected for use in training the model; data labelling, which involves how the data to be used is categorised under useful labels; feature selection, which deals with which of the characteristics of the data instances are useful for classification; and classification itself, which is concerned with how the system arrives at a conclusion on which category a data instance would fall under. The report below is divided

Grenze ID: 01.GIJET.9.2.559

into five sections: the first is the problem statement which acutely defines the research question that the paper is intended to answer, the second is related work, which offers a summary of the literature reviewed in the research process, the third is a problem analysis, which is an in-depth description of the problems discovered during research, the fourth deals with proposed solutions and their justifications, which consists of solutions discovered within the perused literature and how viable they are for solving the said problems and the fifth is the conclusion, which summarizes the findings of this research undertaking.

II. Problem Statement

Given the different applications of speech emotion recognition across different disciplines and industries, the question remains to be answered as to whether current machine learning approaches, including algorithms, decision mechanisms and neural networks, provide the accuracy needed for the output of speech emotion recognition systems to be rendered useful. This research paper aims to answer this question through an eventual comparative analysis of the techniques to be elicited further along, and current accuracy rates are found to be insufficient, an attempt would be made by this research paper to propose alternative methods, based on supplementary research.

III. Related Work

Schuller [9] elicits a standard framework for approaching the design of speech emotion recognition engines, comprising 4 components: modelling, annotation, audio features and textual features. At the modelling stage, an emotional representation model needs to be constructed that is both disciplinarily sound, as regards the field of psychology, and intelligible to a computer.

The scope of emotions to be trained and tested by the model, therefore, is typically defined based upon Ekman's 'Big Six' emotion categories and metrics for these emotions are arousal, which deals with the acoustic portion of the input, and valence, which is concerned with the linguistic content of the input.

Features to be analysed in a speech emotion recognition system are categorised into spectral and prosodic features [5]. According to [7], spectral features, which focus on the biological features of sound production are typically measured using the Mel Frequency Cepstral Coefficient (MFCC), which is an anti-noise technique and as such, has proven effective for isolating relevant sounds in speech inputs and processing them and the Linear Perception Cepstral Coefficient (LPCC), which is adept at extracting vowel sounds made by the speaker, which is instrumental in deciphering the words that are spoken. Prosodic features on the other hand, which cover variable characteristics of the voice, are measured using a wider variety of methods. These include the Hidden Markov Model (HMM), the Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) [7].

Nantasri et al. [6] take a critical look at the aforementioned MFCC method for analysing and evaluating spectral features, to optimise its output, given the level of promise it has shown in improving SER system accuracy. They propose an averaging principle which when used with artificial neural networks, preferably recurrent variants, in their literature, has outperformed the accuracy of two systems proposed by two different sets of previous researchers. On the other hand, [2] takes a bird's eye view approach of extractable spectral features, such as MFCC, LPCC and TEO, which stands for Teager Energy Operator, a lesser-known extractable audio feature, and traditional classifiers as well, such as HMM, GMM and Vector Quantization (VQ), dissecting and evaluating each of them, providing a comparative outline of their strengths and deficiencies as necessary. They state emphatically that there is no agreed-upon best classifier for an SER system, however, each can be optimised in conjunction with other methods to improve its performance.

The chronology of SER system development is explicated in [1], with a focus on database selection and model creation. A perspective, different from the ones included in the pieces of literature outlined above is introduced in this paper, where the deep learning architectures are proposed as a viable alternative to the classical methods of SER system development. This is further supported by [4], in which the performance rates of different linear and nonlinear classifiers that are used in traditional SER systems are contrasted with the performances of deep learning networks.

Huang et al. [3] and Xu et al. [10] both document the development of an SER system with a specific objective. The system developed by [3] is aimed at recognising mood disorders in psychiatric patients and through the study of patients' long-term emotional states, helps psychiatrists make accurate and relevant diagnoses. A peculiar issue faced in this piece of literature is the inability of the developers to make use of an existing dataset, because of the non-existence of a dataset containing emotion data for psychiatric patients, which led to the creation of a dataset by the developers, whereas the system developed by [10] was an attempt to improve on the

accuracy of systems implemented in cited pieces of literature, using different methods on two widely known emotion databases: the IEMOCAP and RAVDESS datasets.
Saxena et al. [8] provide a holistic overview of some of the emotion recognition technology that exists in artificial intelligence today. This includes facial emotion recognition (FER), speech emotion recognition (SER), textual emotion recognition (TER) and the use of physiological signals.

IV. Problem Analysis

From the initial conception of an idea to create a speech emotion recognition system in the first place to the completion and deployment of a high-accuracy system, there are various steps undertaken by developers (Fig.1), each of which comes with its own problems that may negatively impact overall accuracy.

Figure 1. Traditional Machine Learning Flow Mechanism

This research report identifies four main problems and explores them as they occur within the system realisation process.

A. Dataset Selection/Creation

As a rule of thumb, before any form of artificial intelligence system is developed, a dataset to train the system's model must be decided upon. Speech Emotion Recognition systems are no different and [1] indicate that there are 3 types of dataset developers can choose from when embarking on SER system development:

- Natural datasets, which are obtained from TV shows, YouTube Videos, etc.
- Semi-natural datasets, where actors read a piece of text in different prompted emotional states
- Simulated (synthesised) datasets, where a piece of text is read in different emotional states.

TABLE I. TYPES OF DATASETS

Dataset	Advantages	Disadvantages
Natural Datasets	Authentic, thus reliable	Background noise Existence of concurrent emotions
Semi-Natural Datasets	Similar to natural datasets in authenticity	Actors may suffer the Hawthorne effect (be adversely affected by the idea of being watched) A limited range of emotions is displayed.
Simulated Datasets	A good range of emotions is displayed	Overfitting may occur, as displays of emotion, since being acted out, may differ from real-life emotional expression.

As indicated in Table I, each type of dataset comes with its own advantages and disadvantages which would inevitably impact the learning of the data model and thus, its ability to classify emotions when put into use. The problem arises when it comes to the selection of the type of dataset to be used, in that, any option that is selected would result in a trade-off of some kind; prioritising the merits of one type, would mean forgoing the benefits of the others, and therein lies the developer's dilemma. It is also worth noting that datasets are not always selected, they are also created as shown in [3], depending on the nature of the recognition task the system is expected to undertake. In such cases, everything discussed prior in this section is applicable.

B. Data Labelling

Due to there being a standard procedure for creating an emotion representation model as indicated above in the 'Related Work' section [9], not many obstacles are encountered at the modelling stage. However, speech emotion recognition falls under classification tasks, which make use of supervised machine learning techniques, where the data to be used to train the data model is pre-labelled, and where labelling is ordinarily informed by the existing knowledge of developers or survey participants.
A problem arises, nevertheless, because emotions tend to be subjectively perceived when being expressed. To counteract this, developers have purported many methods of emotion identification by humans for labelling, all of which have their own limitations. The first of which will be discussed is acting, where the person whose emotional response is being sampled for model training is prompted with emotion and they are required to act out their interpretation of the said prompt in speech, which would then be recorded. The second is arbitrary

identification, where a third party is prompted the recording another participant's emotional expression, and they are required to identify the emotion being displayed. The former has the limitation of ingenuity, where since the participant is not truly experiencing the expressed emotion, the enactment may not be representative of the appropriate emotion and may mislead the system in its classification, whereas the latter's deficiency lies in the subjective interpretation of emotion where one's interpretation of a given expression may differ from the next.

C. Feature Selection

The third problem to be addressed to answer the above research question is determining which features, whether spectral, prosodic or otherwise, contain the most emotionally relevant content to optimise the data model's processing of inputs [5]. Schuller [9] indicates that spectral features while 'low-level' are robust to external influence, such as background noise, whereas prosodic features, such as intensity, rhythm and intonation are more telling of a user's emotional state due to the possibility of extreme values.

D. Classification Methods

Lastly, there is the dilemma of deciding which classifier is most appropriate for developing optimally accurate speech emotion recognition systems. This involves choosing from a range of linear and nonlinear classifiers. Linear classifiers are classification mechanisms in machine learning that separate or classify a range of data points into discrete sets separated by a line. These include Bayesian Networks, the Maximum Likelihood Principle and the Support Vector Machine. On the other hand, nonlinear classifiers are designed to perform classification tasks on data instances that are "linearly inseparable". This classifier type includes the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM) [7]. Typically, artificial neural networks are used in conjunction with the above classification mechanisms implemented as components, as they tend to simulate the human brain's processing of information.

V. Suggested Solutions for SER Problems

A. Dataset Selection/Creation

The decision as to whether to create a dataset or to choose an existing one to train a model is solely dependent on the type of Speech Emotion Recognition task being undertaken. If generic SER is being pursued, an existing dataset would be more appropriate, as shown in [10] with the IEMOCAP and RAVDESS datasets. However, a niche SER task, such as an SER system for psychiatric patient diagnoses [3] may require a dataset that is more representative of the type of emotional responses the system is expected to encounter.

Secondly, the type of dataset to be selected or created is determined by the priorities of the developers of the system. As indicated in the problem analysis, choosing one type of dataset implies forgoing the benefits of the others.

In the event, however, that best efforts to obtain an adequately representative dataset still yield an imbalanced distribution, the unweighted accuracy of the dataset, which refers to the mean accuracy of each emotion class, can be evaluated and fed into the data model, supplemented by the weighted accuracy of the dataset, which is the mean accuracy of all emotion instances. This will offset any imbalances.

B. Data Labelling

Granted that acting and arbitrary identification, the data labelling methods listed above, both have their unique strengths and deficiencies, a triangulation approach can be adopted in an attempt to neutralise the impact of the deficiencies and underscore the strengths of both methods [9]. This would mean that two datasets or however many are necessary of both types could be combined into a single dataset. Alternatively, labelled data obtained according to the first method could also be subjected to arbitrary identification, to obtain a second opinion on acted-out emotions, which could be compared with the labels obtained from the acting, from which an average can be reached for each emotion data instance.

C. Feature Selection

The two most ubiquitous spectral features are Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) [2]. The former is a representation of a speech signal where the cepstrum is derived from a windowed time frame from that signal. It uses the Mel scale, which mimics the frequency response of the human ear, which makes it apt for emotion detection. The latter, on the other hand, is a digital method for encoding an analogue audio signal, and its function is to predict the next value of an aspect of a signal using previously received information from earlier portions of that signal [2].

While [6] proposes a system that makes sole use of MFCC feature extraction, [2] asserts that no spectral feature should be used on its own to inform an SER system. This is because they each have deficiencies, which when used on their own, would cause a system to be blindsided by important factors which could adversely affect a classification decision. Thus, a triangulation approach is recommended.
Prosodic features, while not as expository as spectral features, are still included in the feature extraction process [1]. This is because these features, for example, fundamental frequency, duration, speaking rate and intensity, can be very telling of an emotional state and can greatly supplement the findings of spectral feature extraction.
Lastly, speech content awareness is crucial [1]. This is because, despite triangulation efforts, there is still the possibility of the odd occurrence that deductions based on features may conflict with each other. Thus, knowledge of the speech content may provide clarity wherever such confusions arise.

D. Classification Methods

Despite the application of linear classifiers, such as the SVM, in SER systems, [4] indicates that nonlinear classifiers are a more appropriate choice for this type of classification task. This is because speech signals are found to be non-stationary, and thus, when represented on a two-dimensional plane, distributions of data instances which may belong to a given category may be irregularly spread across the plane, rendering them linearly inseparable.
While nonlinear classification prevails as the better fit for SER classification, a better classification alternative is proposed in the form of Deep Neural Networks in Deep Learning [1][4]. The advantages that this technique holds over traditional machine learning approaches are given in Table II below.

TABLE II. DEEP LEARNING VS MACHINE LEARNING

Deep Learning	Machine Learning
Neural networks are capable of detecting and quantifying complex features without the need for manual feature extraction and tuning	Manual feature extraction and tuning are required
Networks can deal with unlabelled data, and as such overfitting is not a risk	Networks must always deal with labelled data, as such, overfitting, depending on the weaknesses of the dataset, may be a risk.

Gunawan et al. [2] present evidence from systems that they studied which suggests that the inclusion of DNNs in the construction of an SER system improves the accuracy of the system. The first proposed system that was studied indicated that the inclusion of a DNN in a system provided 20% more accuracy as compared to simple HMM and SVM classification. This is supported by [1] which asserts that the inclusion of deep neural networks such as the Long-Short Term Memory Networks and Deep Convolutional Neural Networks takes the accuracy of a system from values in the seventieth percentile to values in the ninetieth percentile. A comparative analysis by [4] between different traditional classifiers and a deep convolutional neural network is shown in Table III, which makes a deep convolutional neural network stand out.

TABLE III. COMPARATIVE ANALYSIS [4]

Algorithms	Anger	Happy	Sad
k-nearest neighbour	93%	55%	77%
Linear discriminant analysis	68%	49%	72%
Support vector machine	74%	70%	93%
Regularized discriminant analysis	83%	73%	97%
Deep convolutional neural network	99%	99%	96%

Aggarwal et al. [12] also perform a comparative analysis of six algorithms on two emotion speech datasets, RAVDESS and TESS, firstly of the machine learning variety, namely Decision Tree and Random Forest, and secondly of the deep learning variety, namely, Multilayer Perceptron, ResNet 18, and two custom convolutional neural networks, labelled Proposed-I and Proposed-II.
Table IV above summarises the results obtained from the experimentation. The first two models, being traditional machine learning algorithms, yielded relatively low accuracy scores for both datasets. The fourth,

TABLE IV. COMPARATIVE ANALYSIS [12]

Serial Number	Model	RAVDESS	TESS
01	Decision Tree	37.85%	3.21%
02	Random Forest	46.88%	7.68%
03	MLP Classifier	33.68%	15.54%
04	ResNet 18	79.16%	96.26%
05	Proposed-I	73.95%	99.99%
06	Proposed-II	81.94%	97.15%

fifth and sixth models yielded relatively high scores, with the fifth model yielding an accuracy as high as 99.99% for the TESS dataset. The third model, however, stands as an exception, given that it is a deep learning model, yet was unable to yield scores comparable to those of the other deep learning models used in the experimentation.

In addition to the comparisons conducted above, [11] also makes an argument in favour of using deep learning networks, indicating that the temporal and computational cost associated with using deep learning networks for speech emotion recognition tasks is considerably less than that of machine learning algorithms. The literature also asserts that properly selected deep learning networks can identify and learn from features that the developer may even be unaware of, making the learning process more holistic and increasing performance metric scores.

VI. CONCLUSION

Developing an SER system is an intricate and arduous task, consisting of many steps and considerations. Developers would benefit from a thorough understanding of the speech emotion recognition task they are attempting to undertake, in terms of whether it is generic or technically specific and the different types of the dataset available for selection based on their task, feature extraction, selection and submission to the system would have to involve triangulation of different feature types. Above all, however, the evidence points to the fact that the inclusion of deep neural networks in SER systems will give them a sharper edge over traditional machine learning approaches, due to the evident range of advantages associated with their use.

REFERENCES

[1] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models" *Sensors*, 21(4), pp. 1249-1275, 2021.

[2] T.S. Gunawan, M. F. Alghifari, M. A. Morshidi, and M. Kartiwi, "A Review on Emotion Recognition Algorithms Using Speech Analysis", *Indonesian Journal of Electrical Engineering and Informatics* (IJEEI), vol. 6, issue. 1, pp.12-20, 2018.

[3] K.Y. Huang, C. H. Wu, M. H. Su, and Y. T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model". *IEEE Transactions on Affective Computing*, vol.11, issue 4, pp.393-404. 2018

[4] R.A. Khalil, E. Jones, M. I. Babar, T. Jan, M. A. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review". *IEEE Access*, 7, pp. 117327-117345, 2019.

[5] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks. International Journal of Speech and Technology", *International Journal of Speech Technology*, vol. 22, issue 3, pp.497–510, Sep 2019.

[6] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun, A. Rugchatjaroen, and T. Shinozaki, "A Light-Weight Artificial Neural Network for Speech Emotion Recognition using MFCCs and their Derivatives". *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 41-44, 2020.

[7] G. N. Peerzade, R. R. Deshmukh, and S.D. Waghmare, "A Review: Speech Emotion Recognition", *International Journal of Computer Sciences and Engineering*, vol. 6, issue 3, pp. 400-402, 2018.

[8] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems", vol. 2, no. 1, pp. 53-79, 2020.

[9] B. W. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks and Ongoing Trends". *Communications of the ACM*, vol. 61, no.5, pp.90-99, May 2018.

[10] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset". *IEEE Access*, 9, pp.74539-74549, 2021.

[11] Y. B. Singh, and S. Goel, “A systematic literature review of speech emotion recognition approaches”. *Neurocomputing*, vol. 492, pp. 245-263, 2022.
[12] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H.-N. Lee, “Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning”. *Sensors*, 22, p. 2378, 2022.