Emotional Speaker Recognition based on Machine and Deep Learning

Tshephisho Joseph Sefara
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

Tumisho Billson Mokgonyane
Department of Computer Science
University of Limpopo
Polokwane, South Africa
mokgonyanetb@gmail.com

***Abstract*—Speaker recognition is a method which recognise a speaker from characteristics of a voice. Speaker recognition technologies have been widely used in many domains. Most speaker recognition systems have been trained on normal clean recordings, however the performance of these speaker recognition systems tends to degrade when recognising speech which has emotions. This paper presents an emotional speaker recognition system trained using machine and deep learning algorithms using time, frequency and spectral features on emotional speech database acquired from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). We trained and compared the performance of five machine learning models (Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and k-Nearest Neighbor), and three deep learning models (Long Short-Term Memory network, Multilayer Perceptron, and Convolutional Neural Network). After the evaluation of the models, the deep neural networks showed good performance compared to machine learning models by attaining the highest accuracy of 92% outperforming the state-of-the-art models in emotional speaker detection from speech signals.**

Keywords— RAVDESS, neural networks, machine learning, emotion recognition, speaker recognition

I. INTRODUCTION

Deep learning is a machine learning technique that imitate the functions of a human brain in processing structured and unstructured data for intelligent applications such as machine translation, speech recognition, object detection, and many more. Deep learning techniques can learn and be able to make decisions without supervision of a human being. Thus, these techniques have attracted significant research interests in image recognition, computer vision and natural language processing. Famous commercial companies such as Apple, IBM, Facebook, Oracle, Amazon, Microsoft, Google, etc. use deep learning approaches to help scale their business rules. Deep learning techniques are the current more successful approaches in Data Science competitions which were previously dominated by other machine learning techniques such as support vector machine (SVM). Deep learning techniques have illustrated successful results in emotion and speaker recognition over traditional approaches [1]–[5].

Speaker recognition can be defined as a method that recognises a user from speech. The performance of speaker recognition is affected by many factors, such as background noise, channel effect, speaker, and quality of the recordings. These factors may bring negative influence on speaker recognition by inducing extra intra-speaker vocal variability, which is the difference across speakers.

Most speaker recognition systems lack quality due to low amount of training data size for target language. This environment is called low-resourced [32]. Mokgonyane et al. [31] proposed a speaker recognition system that utilizes machine learning models. The models are trained on a clean speech database of low-resourced language. Authors obtained good accuracy of 96% using neural networks. Though the data did not contain emotional speech. Emotion is another internal source which can induce intra-speaker vocal variability [6]. Emotion recognition is a technique used to recognise emotion. In many cases, a speaker recognition system is built on a normal speech data set and this system degrade in performance when tested on speech emotions data set. This paper proposes a speaker recognition system trained using speech emotion data set consisting of eight emotions (disgust, surprise, fearful, angry, sad, happy, calm and neutral), and thus referred to as emotional speaker recognition system. Three sets of acoustic features of speech (Time, Frequency, and Spectral) are extracted from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data set and used to train and compare eight different learning models. We list the following contributions as follows:

- Speaker and emotion recognition literature review is presented.
- We provide list of features used to train the models.
- We train speaker recognition using emotional database and obtain good performance.
- We provide the best SVM kernel for speaker recognition.
- We analyze the performance of deep learning approach and machine learning techniques.
- We suggest models to predict speaker emotions.
- We provide how models distinguish male and female speakers.

This paper is outlined as follows. Firstly, the literature review study on speaker and emotion recognition is discussed in Section II and Section III explains the data processing, feature engineering (extraction and normalization), and methods used to build and evaluate the models. The discussions and findings are given in Section IV, while Section V discusses the conclusion of the paper.

II. Literature Review

This section explains the literature review based on current speaker recognition systems and emotion recognition systems

A. Speaker Recognition

The performance of a speaker recognition can be affected by factors such as quality of the speech, age of the speaker, gender of the speaker, background noise of the recording, accent of the speaker, and other factors. Mbogho and Katz [7] investigated how accent can affect the quality of speech recognition by training two types of Hidden Markov models (HMM). Authors trained the first model on native speakers of English and the second model on accented English speakers. The results showed that models trained on accented English speakers performed better compared to the model trained on native speakers. There are various ways the quality of a speaker recognition system can be measured and according to Ferrer et al. [8], the performance scores of the speaker recognition system can be difficult to interpret. Hence, Ferrer et al. [8] proposes a trial-based normalisation technique to apply on the performance scores to transform them into actual likelihood proportions, that may have a perfect probabilistic interpretation. When evaluating speaker recognition system, Wu et al. [9] shows that data dependence can affect the standard error of the cost function. Neural networks played a significant role in training speaker recognition models. A famous deep learning method called dropout has demonstrated significant improvement in the performance of complex neural networks [10]. The use of neural networks in speech recognition has motivated Tang et al. [1] to use multitask recurrent neural network models to propose collaborative joint training approach for speech and speaker recognition systems.

There are different features that can be extracted from speech to train a speaker recognition model. One of the feature representation is an i-vector, that models both the speaker and channel variability provided in speech signal. Xu et al. [11] propose an approach to extract i-vector without evaluating the full posterior covariance by accelerating the extraction process at run-time. This is attained by generalising the estimation of i-vector, while Cumani and Laface [12] propose e-vector, a speaker modelling technique which generates a compact representation of a speech segment, similar to i-vectors. Modipa et al. [13] investigated different techniques to the acoustic modeling of under-resourced language, Sepedi, for speech recognition while Manamela et al. [14] create an emotion recognition system for the same language using machine learning algorithms.

B. Emotion Recognition

Emotion recognition is the identification of emotions in a rendered speech signal. There are various methods exist to train emotion recognition systems. Deep neural networks are among famous models to train emotion recognition systems. Even though such models are expensive to train, a light deep neural network model for recognition of emotions in audio-visual is proposed by Vielzeuf et al.[2]. Authors reported to have obtained a state-of-the-art accuracy of 61%. Attention has improved the performance of deep neural network due to its technique to attend to more relevant features that predicts the target. As proposed by Ma [3], a multi-task attention-based deep neural network on speech emotion recognition outperforms random forest, deep neural network, and SVM techniques. Egorow et al. [15] use feature selection based on random-forest technique to select the most important features and achieved an increase in performance using only 40 to 60% of the features using *emobase* feature set [16]. Marczewski et al. [17] propose a hybrid speech emotion recognition architecture where the first layer uses a convolutional neural network (CNN) as a feature extraction step and final layer is a classification layer that consists of a long short-term memory network (LSTM), for emotion classification using domain-specific features. Sun et al. [4] use CNN-LSTM to extract features of film characters and SVM is used for classification while Wang and Hu [18] use SVM on improved Mel Frequency Cepstral Coefficients (MFCCs) to obtain state-of-the-art results. Albanie et al. [4] proposes a technique of labelling unlabelled speech emotion dataset by using a pretrained emotion recognition neural network trained on images. Studies show that noise has negative impact on speaker and emotion recognition systems and thus Pohjalainen et al. [19] propose signal denoising in log-spectral and cepstral domains and authors ascertain that the method proposed performed better than conventional noise reduction methods.

III. Methodology

This section firstly discusses the acquired data, secondly discusses the feature extraction and normalisation techniques, thirdly explains the models, and lastly discusses how the models are evaluated.

The proposed emotional speaker recognition system architecture is depicted in Figure 1. The first step is to extract acoustic features from the speech database, then the features are normalised using z-score feature normalization. Lastly, we use the training data to build the models and the testing data is used to evaluate the models by making predictions and compare with true label.

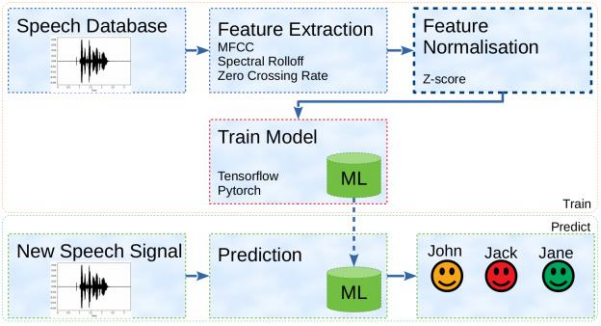

Fig. 1. Overall system overflow

A. Data

This study uses a validated emotional speech and song multimodal database, RAVDESS [20], collected from 24 professional speakers (12 females and 12 males), recorded the same speech in a neutral North American accent. Utterances are composed of eight emotions: disgust, neutral, surprise, calm, fearful, angry, happy, and sad expressions. We removed the silence in the speech signals. Then we resampled speech signals to 32000 samples per second and Table I shows the total duration of each speaker for all the recorded emotions. The paper does not include the songs.

TABLE I. SPEAKER DURATION FOR 24 SPEAKERS

Speaker	*Duration*	*Speaker*	*Duration*
01	03:44	13	03:19
02	03:47	14	03:40
03	03:45	15	03:29
04	03:37	16	03:43
05	03.44	17	03:40
06	03:47	18	03:44
07	03:44	19	03:52
08	03:43	20	03:43
09	03:29	21	03:55
10	03:44	22	03:43
11	03:26	23	03:36
12	03:45	24	03:57

Fig. 2 illustrate an example of the same linguistic content for the neutral and anger speech signals. The spectrograms in Fig. 2a show most portion of the spectral energy concentrate at lower frequencies in a voiced region of the neutral speech (in most cases below 512 Hz). The spectral energy is spread over a range of frequencies and there is no harmonic form in the anger utterance. It can be seen that the anger speech has amplitude of around ±1 while neutral speech has a lower amplitude of around ±0.04. Thus for English language, people express emotions in speech differently.

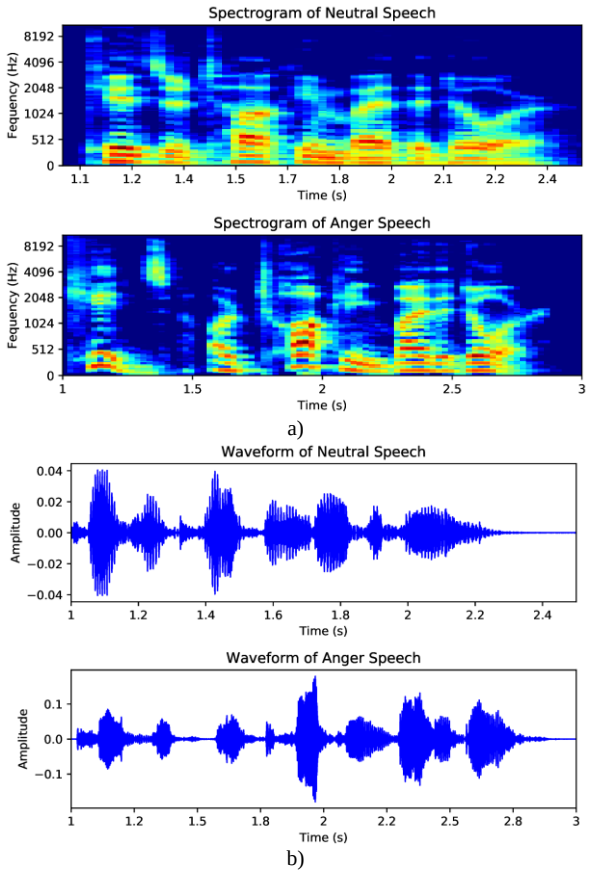

Fig. 2. Spectrogram and waveform of neutral and anger utterance produced by one speaker.

B. *Feature Extraction and Normalisation*

This section discusses extracted features and type of normalization used.

1) Feature Extraction: The speech utterance has different acoustic features which characterises the identify of the speaker. We use pyAudioAnalysis [21] to extract short-term features shown in Table II to create a feature vector of size 68 consisting of both the standard deviation and mean. During extraction, we set the Hamming window to a rate of 25ms and frame size of 50ms. The 34 short-term extracted features are categorised into 3 domains (Frequency, Time, Cepstral) and the descriptions of the features are given in Table II. These features are also used in [22]-[24]:

- **Time-domain features** include Energy, Zero Crossing Rate (ZCR) and Entropy of Energy. These features are extracted from the speech recordings. Using the average ZCR, a representation is determined and used to calculate the estimates of the spectral properties. A definition of calculations is as follows [25]:

$$Z_n = \sum_{m=-\infty}^{\infty} |\ \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\ |\ w(n-m)$$

$$\begin{aligned} &where \\ &\quad \mathrm{sgn}[x(n)] = 1 \quad x(n) \geq 0 \\ &\quad\qquad\ = -1 \quad x(n) < 0 \\ &and \\ &\quad w(n) = 1/(2N) \quad 0 \leq n \leq N-1 \\ &\quad\qquad\ = 0 \qquad otherwise \end{aligned} \tag{1}$$

- **Frequency-domain features** include Chroma Vector, Chroma Deviation, Spectral Rolloff, Spectral Entropy, Spectral Flux, Spectral Centroid, and Spectral Spread which are based on the size of the Discrete Fourier Transform (DFT).

- **Cepstral-domain features** include MFCCs which are determined when an inverse DFT is applied on the logarithmic spectrum. MFCCs are commonly used as acoustic features of speech in emotion and speaker recognition applications [18], [26]. MFCCs are calculated as follows:

$$c_n = \sum_{m=1}^{M} \left[\log Y(m)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{2}$$

where the index of a cepstral coefficient is represented by *n*, *Y(m)*, *m=1*, ..., *M*, is the output of an M-channel filterbank.

TABLE II
ACOUSTIC FEATURES ON SHORT-TERM WINDOWS [21]

Feature ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	MFCCs form a cepstral representation where the frequency bands are distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

2) Feature Normalisation: is a significant step to create a robust machine learning model for speaker recognition. Normalisation has been used for speaker and emotion recognition systems [27]-[29]. The aim is to remove recording and speaker variability while retaining the afficacy of speaker discrimination. We use z-score normalisation defined by Sefara [22] that is formulated using the following equation:

$$\hat{y} = \frac{S - \bar{x}}{\sigma} \tag{3}$$

where the mean is represented by $\bar{x}$, standard deviation is represented by σ and ŷ is the estimated standardised feature.

C. Models

This section explains the learning models implemented on the RAVDESS data set.

1) Logistic Regression: is a non-linear transformation of the linear regression shown in Fig. 3. The logistic distribution is an S-shaped distribution function. For binary classification, the logit distribution includes probability estimates to fall in (0 - 1). Logistic regression equation takes the form:

$$logit[p(x)] = \log \frac{p(x)}{1 - p(x)} \tag{4}$$

where *logit[p(x)]* is the $\log_e$ of the likelihood ratio that the dependent variable is 1, and *p* ranges in (0 – 1).

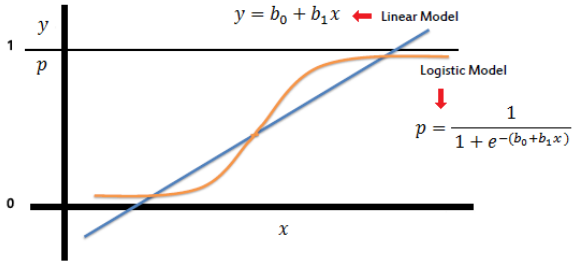

Fig. 3. Linear and Logistic Regression Lines.

2) k-Nearest Neighbors (kNN) is a machine-learning technique which classifies a data point using its k nearest neighbors. kNN has different properties: (i) kNN is non-parametric since it does not assume the probability distribution of the input data point. (ii) kNN use lazy learning method since it generalises during testing not training phase.

3) Random Forest makes classification by creating decision trees on samples of data as shown in Fig. 4 during training and output the category that is the mode of the categories using mojority voting.

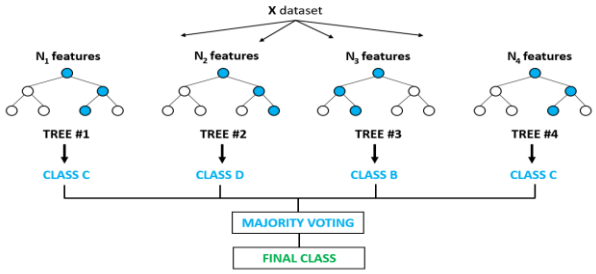

Fig. 4. Random Forest.

4) Extreme Boosting is an optimised implementation of gradient boosted decision trees specifically designed to be fast and efficient.

*5) SVMs a*re machine learning models that have different kernels used for regression and classification problems. SVM is a discriminative classification model that creates an separating hyperplane to categorise a new data point as shown in Fig. 5. The following SVM kernels are implemented.

- Linear SVM: $<x,x'>$
- RBF SVM: $exp\ (-\gamma \| x-x' \|^2)$
- Polynomial SVM: $(\gamma < x,x' > + r)^d$
- Sigmoid SVM: $tanh\ (\gamma < x,x' > + r)$

where *r* is the coefficient, gamma γ is always positive, and *d* is a kernel degree.

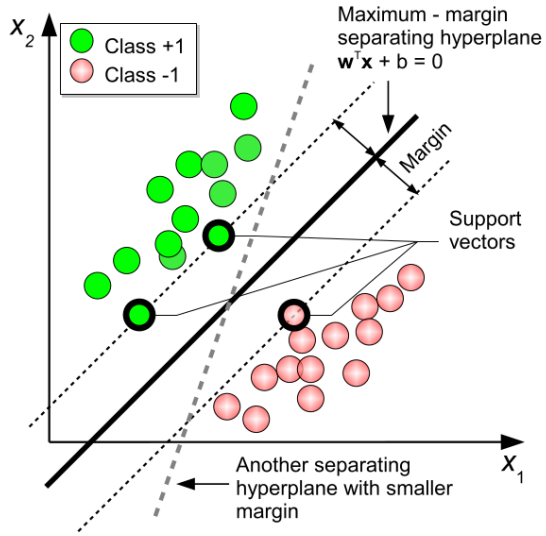

Fig. 5. SVM showing a separating hyperplane [30].

6) Multilayer Perceptron (MLP) is a type of feed-forward neural networks. MLP consists of multiple layers activtated by different activation functions. Table III shows the architecture of the implemented MLP. We use Tensorflow to implement the MLP architecture. The dropout layers helps to prevent overfitting. The model is trained for 1000 iterations using batch size of 128, and contains 31704 parameters. The model accuracy and loss are shown in Fig. 6 and 7 respectively.

TABLE III. MLP ARCHITECTURE

Layer	*Output Shape*	*Parameters*
Dense	(128, 128)	21888
Dropout	(128, 128)	0
Relu	(128, 128)	0
Dense	(128, 64)	8256
Dropout	(128, 64)	0
Relu	(128, 64)	0
Dense	(128, 64)	1560
Softmax	(128, 24)	0

*7) CNNs a*re type of deep neural networks mostly applied in the domain of computer vision. Table IV shows the architecture of the implemented CNN. We use Tensorflow to implement CNN and add dropout regularisation with probability of 0.5 to avoid the model from overfitting. The model is trained for 1000 iterations using batch size of 128, and contains 112024 parameters. The model accuracy and loss are shown in Fig. 6 and 7 respectively.

TABLE IV. CNN ARCHITECTURE

Layer	*Output Shape*	*Parameters*
CNN	(128, 1, 128)	108928
Dropout	(128, 1, 128)	0
Global average pooling	(128, 128)	0
Dense	(128, 24)	3096
Softmax	(128, 24)	0

8) LSTMs are special kind of recurrent neural networks, with capability to learn long-term dependencies. Table V shows the architecture of the implemented LSTM. We use Tensorflow to implement LSTM and add dropout regularisation with probability of 0.5 to avoid the model from overfitting. The model is trained for 1000 iterations using batch size of 128, and contains 156184 parameters. The model accuracy and loss are shown in Fig. 6 and 7 respectively.

TABLE V. LSTM ARCHITECTURE

Layer	*Output Shape*	*Parameters*
LSTM	(128, 128)	153088
Relu	(128, 128)	0
Dense	(128, 24)	3096
Softmax	(128, 24)	0

Table III-V shows the architecture of the models implemented as sequence models which consist of multiple layers where each layer has a shape which represents number of neurons. The last layer contains output shape of 24 which must be equals to the total number of speakers.

IV. EVALUATION

The quality of the model can be affected by number of factors such as the size and the amount of the noise in the training data, the quality of the recorded voice, the type of the recording device, and type of learning technique. In every machine learning pipeline, the evaluation of how the models generalise on unseen data should be considered. From a total number of 1296 speech samples. We splitted the data into 80%, 10% and 10% for training, testing and evaluation respectively. We selected the following list of measurements to measure the prediction quality of the models using the test data and evaluation data:

a) Accuracy: is the fraction of the sum of the true positives and true negatives among all the elements in the test data. Equation to calculate accuracy is determined as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{5}$$

b) Categorical cross-entropy: it is a Cross-Entropy loss plus a Softmax activation, sometimes called Softmax loss. Categorical cross-entropy loss function is a best selection for categorical data. The formulation is determined as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{N} 1_{y_i \in C_c} \log p_{model}[y_i \in C_c] \tag{6}$$

where $i \in [1...N]$ is an observation and $c \in [1...C]$ represents categories. The $p_{model}[y_i \in C_c]$ is the probability forcasted for the element 'i' to belong to the category 'c'.

c) Precision: is the fraction of the true positives among the total number of real positive elements. Precision answers the question what proportion of positive prediction was correct. The formulation is determined as follows:

$$Precision = \frac{tp}{tp + fp} \quad (7)$$

d) Recall: is the true positives divided by relevant elements. Recall answers the question what proportion of true positives was correctly predicted. The formulation is determined as follows:

$$Recall = \frac{tp}{tp + fn} \quad (8)$$

e) F_1 score: is a measure of the accuracy of the model and it considers both the recall and the precision. F_1 score is calculated using the following equation:

$$F_1 score = 2 \times \frac{precision \times recall}{precision + recall} \quad (9)$$

f) Confusion matrix: is a table used to assess the performance of the classifier where true values are known. In the table the diagonal elements (*tn* & *tp*) are the number of examples where the forecasted label is the same as the true label, while off-diagonal elements represent examples that the classifier incorrectly labelled.

V. Findings and Discussions

This section explains the performance and overfitting after training the models using accuracy, F_1 score and categorical cross-entropy.

A. Performance

We show the performance results in Table VI after evaluation of the models. To show the best SVM kernel for speaker recognition using the selected features, we trained SVM on four kernels, namely, sigmoid, linear, RBF, and polynomial. The RBF kernel performed better than polynomial, linear and sigmoid. The Sigmoid SVM obtained poor performance. We notice Sigmoid kernel failed to obtain the state-of-the-art accuracy with 58% being last followed by polynomial, linear and RBF kernels with 81%, 85%, and 88% respectively obtaining the state-of-the-art accuracy. From these results, RBF kernel is suitable for speaker recognition system and sigmoid is not suitable.

We group machine learning algorithms (LR, RF, kNN, XGBoost, SVM) and deep learning algorithms (MLP, CNN, LSTM) to show the best technique for speaker recognition using RAVDESS dataset. For machine learning algorithms, we observe better results by RBF SVM obtaining highest accuracy. For deep learning algorithms, we observe MLPs performing better with 92% followed by LSTM and CNN. Thus deep learning models as shown in Table VI performed better with regard to F_1 score and accuracy than machine learning models, hence, are the best models to use in emotional speaker recognition. We also observe LR obtaining state-of-the-art results outperforming RF, kNN, XGBoost, sigmoid SVM, linear SVM, and polynomial SVM.

Figure 8 shows the confusion matrix after evaluating MLP which in this case, is the best model for speaker recognition on emotional dataset. The even numbers represent female speakers and odd numbers represent male speakers. We observe that MLP confused male speaker with another male speaker for 11 instances and 1 instance for female and male speaker. Although, the model is not trained for gender classification but this could mean the model learned the difference between male and female speakers and which features are shared among same gender.

TABLE VI. Evaluation Results

Model	*Accuracy*	*F_1 score*
LR	0.85	0.86
RF	0.81	0.79
kNN	0.79	0.77
XGBoost	0.83	0.83
Linear SVM	0.85	0.86
RBF SVM	0.88	0.88
Polynomial SVM	0.81	0.81
Sigmoid SVM	0.58	0.58
MLP	**0.92**	**0.92**
CNN	0.89	0.89
LSTM	0.90	0.90

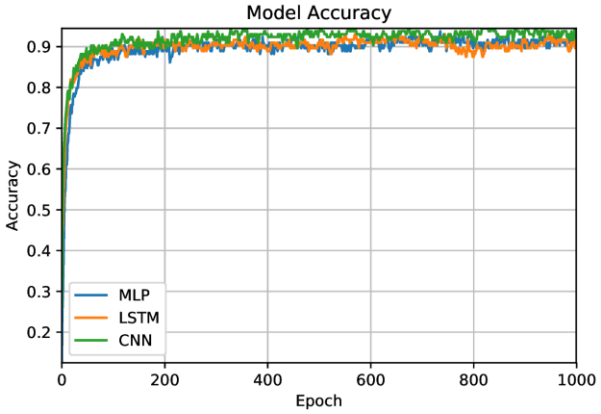

Fig. 6. Accuracy for MLP, LSTM, and CNN

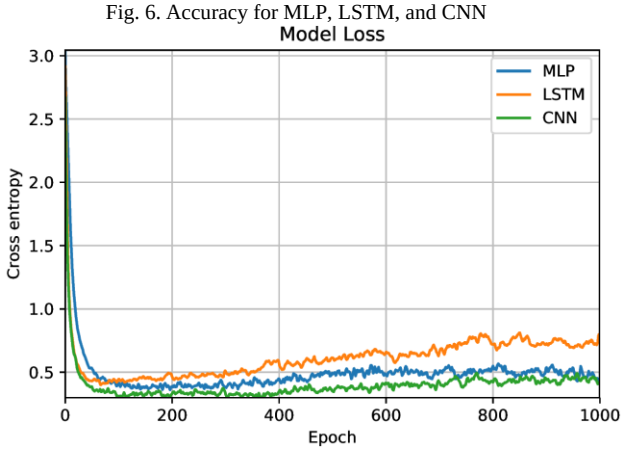

Fig. 7. Categorical cross entropy for MLP, LSTM, and CNN.

B. Overfitting

Overfitting happens when a model learns the training data well to a point where the trained model fails to predict the new data. We investigate overfitting by investigating the model's learning curves shown in Figure 6 for LSTM, MLP, and CNN. We trained the models for 1000 iteration and as shown in Fig. 6 the prediction accuracy did not reduce. Moreover, we show in Fig. 7, the categorical cross-entropy loss function and it stayed under 0.5 for both CNN and MLP but the loss function started to increase after 300 epochs for LSTM. Hence, MLP and CNN did not overfit but LSTM started to overfit after 300 epochs.

VI. Conclusion

This paper presented speaker recognition system using not just normal database but emotional database which has 8 emotions. We presented literature review on speaker and emotion recognition. The features and feature extraction were discussed. A type of normalisation of features was explained. The learning algorithms were explained. We observed RBF kernel being suitable for speaker recognition among other SVM kernels. We observed deep learning algorithms outperforming machine learning algorithms on emotional database of 24 speakers.

In conclusion, we suggest the extension of this work to include (i) the investigation of the selection of the most relevant acoustic features. (ii) Increasing number of speakers and speech samples

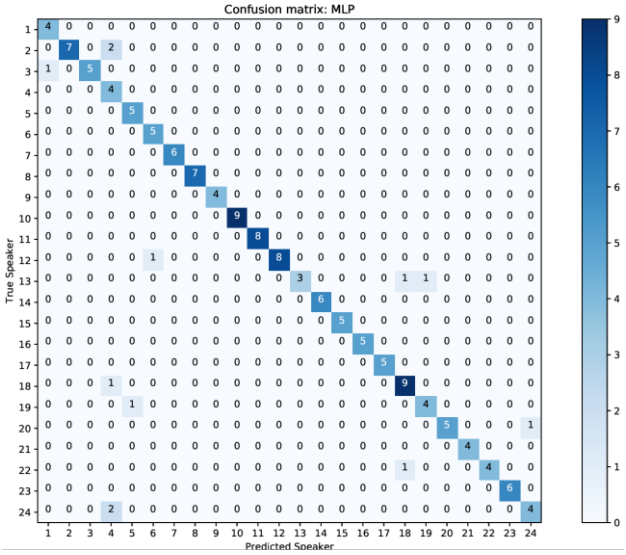

Fig. 8. Confusion matrix for MLP.

References

[1] Z. Tang, L. Li, D. Wang, R. Vipperla, Z. Tang, L. Li, D. Wang, and R. Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 3, pp. 493–504, Mar. 2017.

[2] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie, "An occam's razor view on learning audiovisual emotion recognition with small training sets," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction, ser.* ICMI '18. Boulder, CO, USA: ACM, 2018, pp. 589–593. [Online]. Available: http://doi.acm.org/10.1145/3242969.3264980

[3] F. Ma, W. Gu, W. Zhang, S. Ni, S.-L. Huang, and L. Zhang, "Speech emotion recognition via attention-based DNN from multi-task learning," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, ser.* SenSys '18. Shenzhen, China: ACM, 2018, pp. 363–364.

[4] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "LSTM for dynamic emotion and group emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ser.* ICMI '16. Tokyo, Japan: ACM, 2016, pp. 451–457.

[5] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM International Conference on Multimedia, ser.* MM '18. Seoul, Republic of Korea: ACM, 2018, pp. 292–301.

[6] J. H. Hansen and H. Bořil, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," *Speech Communication*, vol. 101, pp. 94–108, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639317303849

[7] A. Mbogho and M. Katz, "The impact of accents on automatic recognition of South African English speech: A preliminary investigation," in Proceedings of the 2010 Annual Research Conference of the SouthAfrican Institute of Computer Scientists and Information Technologists, ser. SAICSIT '10. Bela Bela, South Africa: ACM, 2010, pp. 187–192.

[8] L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: Trial-based calibration with a reject option," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 1, pp. 140–153, Jan. 2019.

[9] J. C. Wu, A. F. Martin, C. S. Greenberg, R. N. Kacker, J. C. Wu, A. F. Martin, C. S. Greenberg, R. N. Kacker, C. S. Greenberg, J. C. Wu, A. F. Martin, and R. N. Kacker, "The impact of data dependence on speaker recognition evaluation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 1, pp. 5–18, Jan. 2017.

[10] N. Watt and M. C. du Plessis, "Dropout algorithms for recurrent neural networks," in *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, ser. SAICSIT '18. Port Elizabeth, South Africa: ACM, 2018, pp. 72–78.

[11] L. Xu, K. A. Lee, H. Li, and Z. Yang, "Generalizing i-vector estimation for rapid speaker recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 4, pp. 749–759, Apr. 2018.

[12] S. Cumani and P. Laface, "Speaker recognition using e–vectors," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 4, pp.736–748, Apr. 2018.

[13] T. Modipa, M. Davel, and F. de Wet, "Acoustic modelling of Sepedi affricates for ASR," in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, ser. SAICSIT '10. Bela Bela, South Africa: ACM, 2010, pp. 394–398.

[14] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara, and T. B. Mokgonyane, "The automatic recognition of Sepedi speech emotions based on machine learning algorithms," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD).* Durban, South Africa: IEEE, Aug 2018, pp. 1–7.

[15] O. Egorow, I. Siegert, and A. Wendemuth, "Improving emotion recognition performance by random forest-based feature selection," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Leipzig, Germany: Springer International Publishing, 2018, pp. 134–144.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR - introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, Netherlands: IEEE, Sep. 2009, pp. 1–6.

[17] A. Marczewski, A. Veloso, and N. Ziviani, "Learning transferable features for speech emotion recognition," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, ser. Thematic Workshops '17. Mountain View, California, USA: ACM, 2017, pp. 529–536.

[18] Y. Wang and W. Hu, "Speech emotion recognition based on improved MFCC," in *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, ser. CSAE '18. Hohhot, China: ACM, 2018, pp. 88:1–88:7.

[19] J. Pohjalainen, F. Fabien Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. Amsterdam, The Netherlands: ACM, 2016, pp. 670–674. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967306

[20] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, 2018.

[21] T. Giannakopoulos, "pyAudioAnalysis: An open-source Python library for audio signal analysis," *PloS one*, vol. 10, no. 12, pp. 1–17, 2015.

[22] T. J. Sefara, “The effects of normalisation methods on speech emotion recognition,” in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2019, pp. 1–8.

[23] T. J. Sefara and A. Modupe, “Yorùbá gender recognition from speech using neural networks,” in *2019 6th International Conference on Soft Computing Machine Intelligence (ISCMI)*, 2019, pp. 50–55.

[24] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela, and T. I. Modipa, “The effects of data size on text-independent automatic speaker identification system,” in 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2019, pp. 1–6.

[25] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, “Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy,” in *Advanced Techniques in Computing Sciences and Software Engineering*, K. Elleithy, Ed. Dordrecht: Springer Netherlands, 2010, pp. 279–282.

[26] L. Zhao, Q. Wang, and X. Dang, “Recognition influence of different acoustic characters between male and female speakers,” in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, ser. CSAI ’18. Shenzhen, China: ACM, 2018, pp. 394–398.

[27] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, “Speech emotion recognition using Fourier parameters,” *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[28] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai, and P. Lamsrichan, “Speech emotion recognition using convolutional long short-term memory neural network and support vector machines,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Kuala Lumpur, Malaysia: IEEE, Dec 2017, pp. 1744–1749.

[29] A. Tawari and M. M. Trivedi, “Speech emotion analysis: Exploring the role of context,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, Oct 2010.

[30] S. Cumani and P. Laface, “Large-scale training of pairwise support vector machines for speaker recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, Nov 2014.

[31] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa and M. J. Manamela, “Automatic Speaker Recognition System based on Optimised Machine Learning Algorithms,” *2019 IEEE AFRICON*, Accra, Ghana, 2019, pp. 1-7, doi: 10.1109/AFRICON46755.2019.9133823.

[32] T. J. Sefara, “The development of an automatic pronunciation assistant,” M.S. Thesis, Faculty of Science and Agriculture, University of Limpopo, South Africa, 2019. [Online]. Available: http://ulspace.ul.ac.za/handle/10386/2906